LINEAIRE REGRESSIE: AANNAMES EN CONTROLES

21 mei 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)

VANDAAG

1. Regressieassumpties

...

DE KOMENDE WEKEN

Bijeenkomst	Onderwerp
Dinsdag 14 mei	Lineaire regressie: de basis
Dinsdag 21 mei	Lineaire regressie vervolg: assumpties en controleren
Donderdag 30 mei	Interacties en dummy-variabelen
Dinsdag 4 juni	Logistische Regressie
Dinsdag 11 juni	Multilevel-analyse

TERUGBLIK VORIGE WEEK

LINEAIRE REGRESSIE

Maakt een (lineair) model

om

de waarden te voorspellen

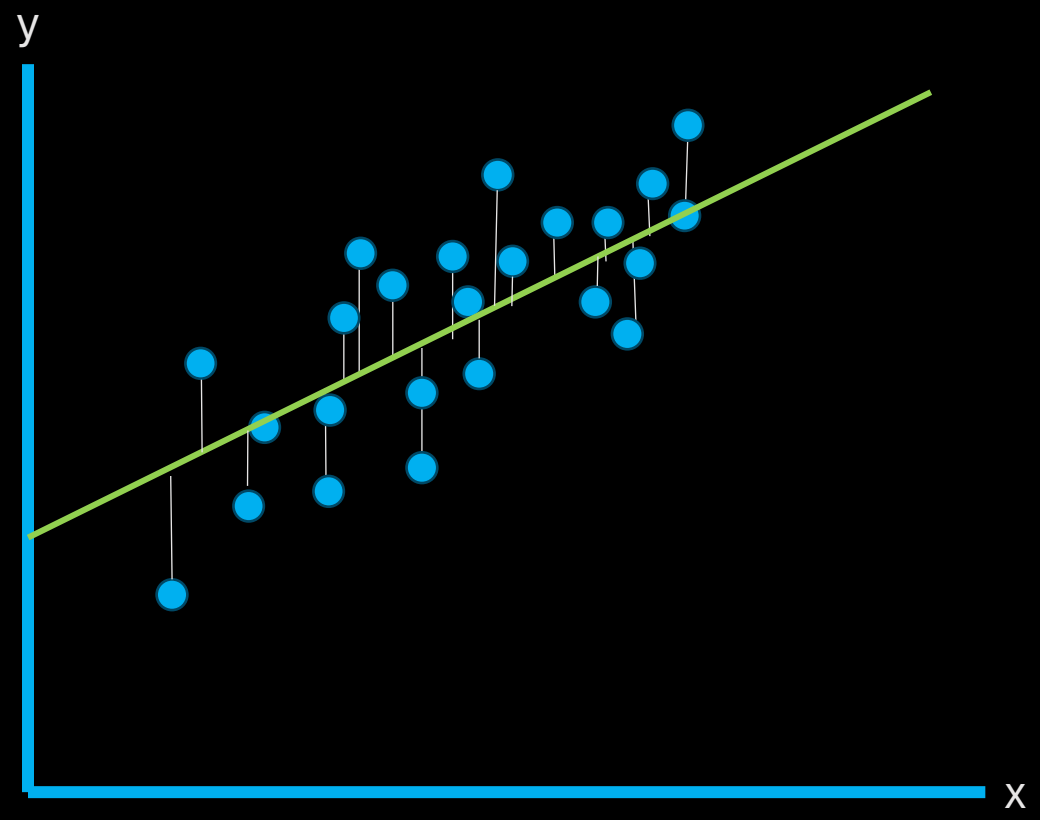
van een *afhankelijke variable* y

met de waarden van

één of meer

onafhankelijke variabelen x

LINEAIRE REGRESSIE IN GRAFIEK



Lijn waarvoor geldt: (gekwadrateerde) som van de verticale afstanden tussen de punten en de lijn is minimaal

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

LINEAIRE REGRESSIE IN VERGELIJKING

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Afhankelijke variabele (waarde voor eenheid i)

Onafhankelijke variabele (waarde voor eenheid i)

Een onverklaarde afwijking (residu) (voor eenheid i) ☹

LINEAIRE REGRESSIE IN VERGELIJKING

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Schatting voor coëfficienten β_0 en β_1
zo dat som van (kwadraat van) de residuen (ε_i)
zo klein mogelijk is

MEERVOUDIGE REGRESSIE

Meervoudige Regressie

Regressie met meerdere onafhankelijke variabelen

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \varepsilon_i$$

Interpretatie van (coefficient) β_0

Als x_1 en x_2 gelijk zijn aan 0, is y gelijk aan β_0

Interpretatie van (coefficient) β_1

Als x_1 omhoog gaat met 1 **en x_2 gelijk blijft**, stijgt y met β_1

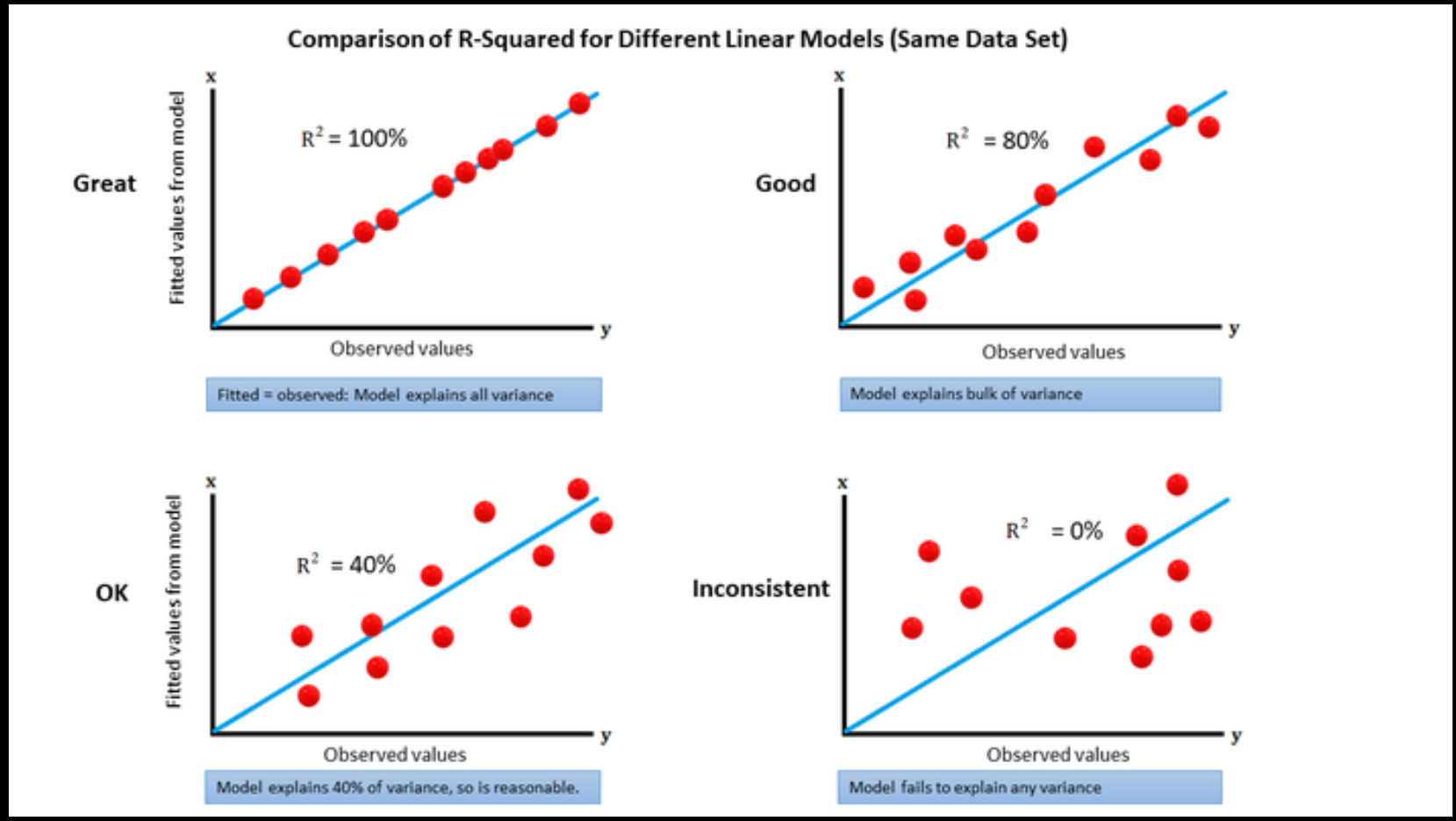
Interpretatie van (coefficient) β_2

Als x_2 omhoog gaat met 1 **en x_1 gelijk blijft**, stijgt y met β_2

Effect is nu *constant houdend voor* andere variabele
Dat gaan we volgende week gebruiken om te *controleren*

DE R^2 (EN ADJUSTED R^2)

De R^2 geeft de ***verklaarde variantie***: een indicatie van hoe goed de gemaakte vergelijking (het "model") de afhankelijke variabele voorspelt.



STANDAARDFOUT

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35595 -0.12057 -0.06430  0.07108  1.81900

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.34603    0.05040  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.02729  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

Formeel: de geschatte standaardafwijking van de *steekproevenverdeling* van de geschatte parameter

Intuïtief: een indicatie van hoe ver we denken dat de schatting (gemiddeld) van de echte waarde af zit.

We verwachten dus dat het effect van hh_grootte 0.48 is, maar met deze steekproef zitten we daar gemiddeld 0.03 naast.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-,346	,050		-6,866	<.001	-,445	-,247
	Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)	,479	,027	,657	17,554	<.001	,425	,533

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

T-TOETS (SIGNIFICANTIE VAN COËFFICIËNTEN)

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35595 -0.12057 -0.06430  0.07108  1.81900

Coefficients:
            Estimate Std. Erro t value Pr(>|t|)
(Intercept) -0.34603    0.0504  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.0272  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

Toets of de gevonden steekproef waarschijnlijk is als de coëfficiënt in de populatie eigenlijk 0 is.

Met andere woorden:
zou je dit effect toevallig kunnen vinden in een steekproef, als er eigenlijk geen effect is.

p < 0.01 voor het effect van hh_grootte:

Het effect is significant: het is niet waarschijnlijk om deze steekproef te vinden als er in de populatie geen effect is.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-,346	,050		-6,866	<.001	-,445	-,247
	Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)	,479	,027	,657	17,554	<.001	,425	,533

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

Let op: de significantie van de constante is inhoudelijk niet interessant

F-TOETS (SIGNIFICANTIE VAN MODEL)

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35595 -0.12057 -0.06430  0.07108  1.81900

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.34603    0.05040  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.02729  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

Toets of de gevonden verklaringskracht van het model (de R2) waarschijnlijk is als het model in werkelijkheid geen enkele voorspellende kracht heeft.

Met andere woorden:
zou je dit gehele model kunnen vinden in een steekproef, als eigenlijk geen onafhankelijke variabele effect heeft.

$p < 0.01$:

Het model voorspelt significant beter dan geen model

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13,033	1	13,033	308,156	<.001[b]
	Residual	17,128	405	,042		
	Total	30,161	406			

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

b. Predictors: (Constant), Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)

Let op: dit is eigenlijk alleen relevant bij meerdere onafhankelijke variabelen.

VANDAAG:
REGRESSIEASSUMPTIES

8 GEVAREN VAN REGRESSIE

DANGER!!

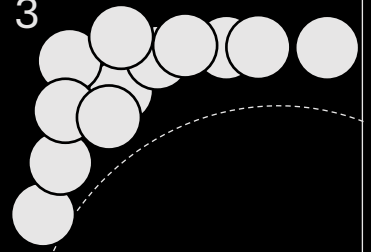
1

Schijnverband

2

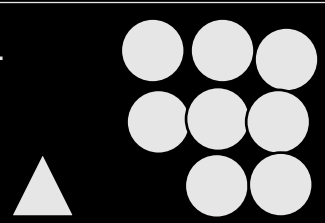
Wederkerigheid / Simultaniteit

3



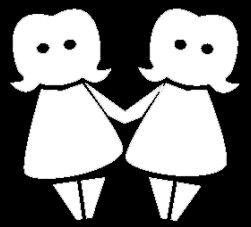
Non-Lineairiteit

4



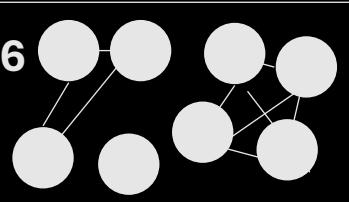
Extreme waarden

5



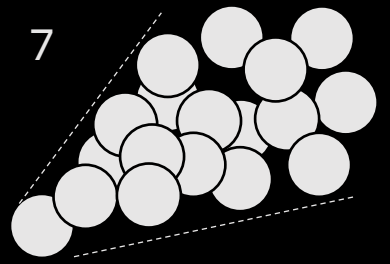
Multicollineariteit

6



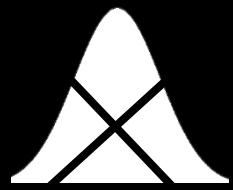
Niet onafhankelijke residuen

7



Heteroskedasticiteit

8



Non-normality of errors

GEVAAR 1

SCHIJNVERBAND / OMITTED VARIABLE / CONFOUNDING VARIABELE

VOORBEELD

Vertrouwen in politici

Deelname in
Demonstraties /
Petities /
Boycotts

RESULTAAT

```
> summary(lm(action ~ trstplt, data = ess10))

Call:
lm(formula = action ~ trstplt, data = ess10)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5009 -0.4203 -0.3720  0.5475  2.6602

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.339765   0.006773   50.16   <2e-16 ***
trstplt     0.016109   0.001514   10.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7221 on 36323 degrees of freedom
  (1286 observations deleted due to missingness)
Multiple R-squared:  0.003108,  Adjusted R-squared:  0.00308
F-statistic: 113.2 on 1 and 36323 DF,  p-value: < 2.2e-16
```

WAT WAS DAT?

Vertrouwen in politici

??????? +

Deelname in
Demonstraties /
Petities /
Boycotts

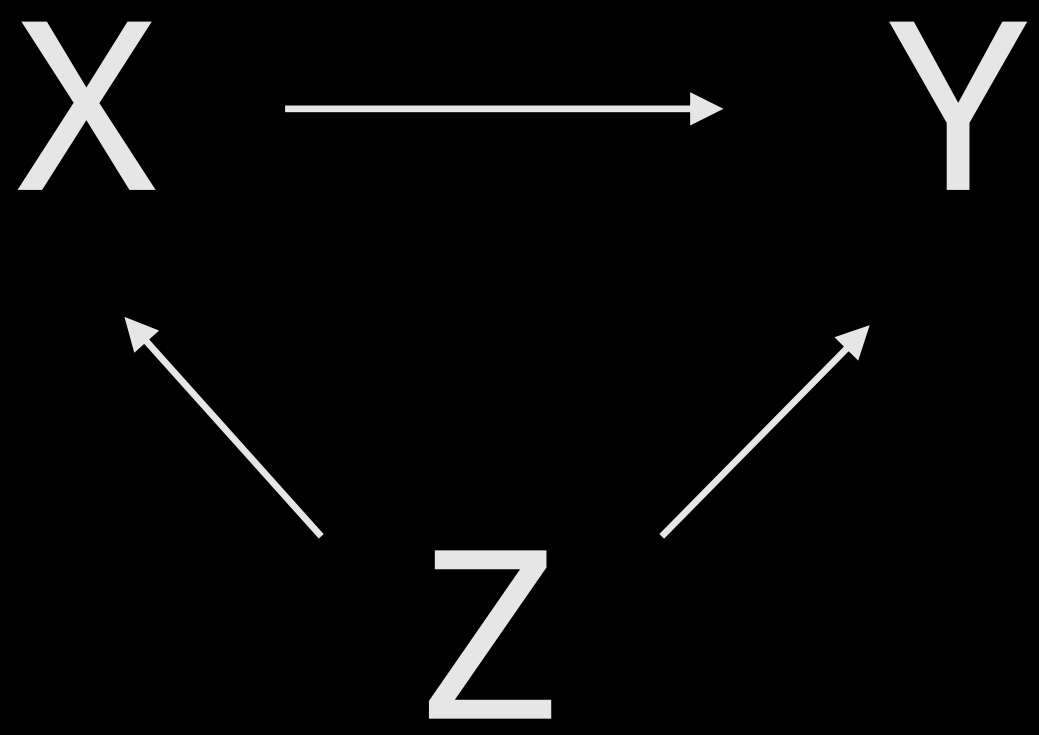
SCHIJNVERBAND

+?

+

+

SCHIJNVERBAND



SCHIJNVERBAND

Een derde **variabele z**
(confounding variable / omitted variable)

heeft een **effect op**
zowel de **afhankelijke variabele**
als de **onafhankelijke variabele**

hierdoor ontstaat er
een **correlatie tussen x en y**
en kan het **ten onrechte *lijken***
alsof er een effect is van x op y

SCHIJNVERBAND

+???

Vertrouwen in politici

Deelname in
Demonstraties /
Petities /
Boycotts

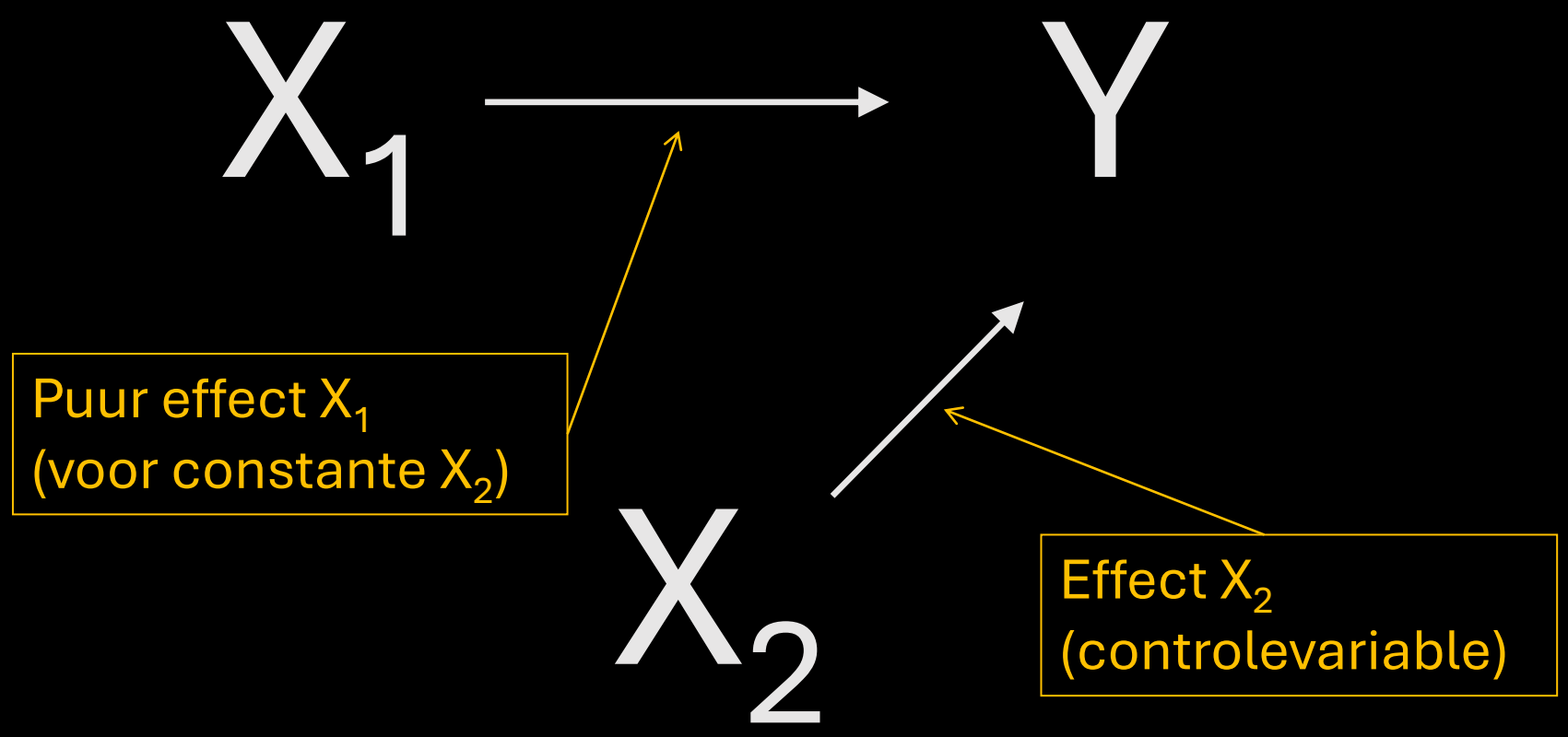
Opleidingsniveau

CONTROLEREN VOOR SCHIJNVERBANDEN

(using multiple regression)

SCHIJNVERBAND: CONTROLEREN

Door de schijnverband-variabele “Z” toe te voegen aan het model kunnen we het probleem verhelpen:
we krijgen het effect **constant houdend voor Z**



GEVAAR 1: SCHIJNVERBAND

Gevaar

Vertekende schatting van coëfficiënten (en dus verkeerde inschatting van de omvang van het effect)

Opsporing

Nadenken. Er is geen (statistische) test die je kan beschermen tegen schijnverband

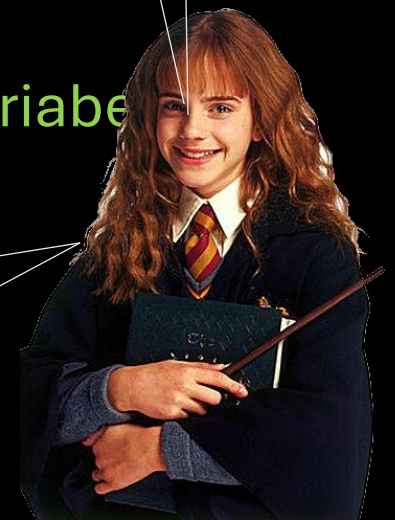
Oplossing

Voeg de *confounding variables* toe als controlevariabe

Effect van X op Y is nu het ‘pure’ effect.

Indien mogelijk…

Als je geen confounding variables vergeten bent…



8 GEVAREN VAN REGRESSIE

DANGER!!

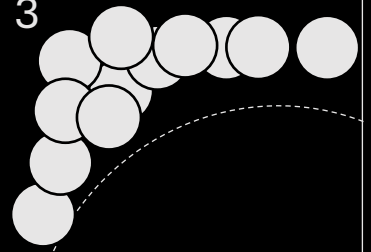
1

Schijnverband

2

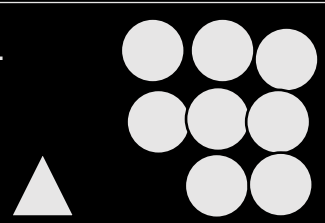
Wederkerigheid / Simultaniteit

3



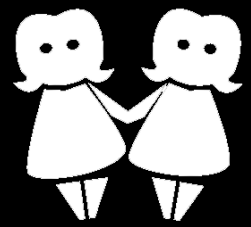
Non-Lineairiteit

4



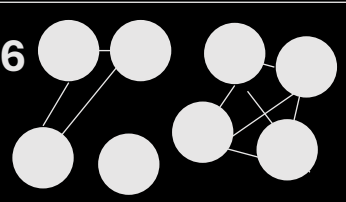
Extreme waarden

5



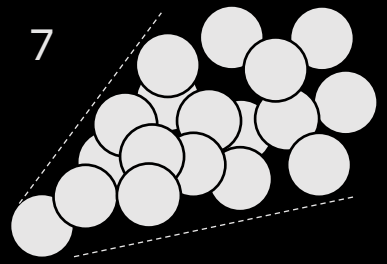
Multicollineariteit

6



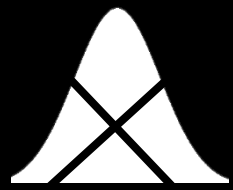
Niet onafhankelijke residuen

7



Heteroskedasticiteit

8

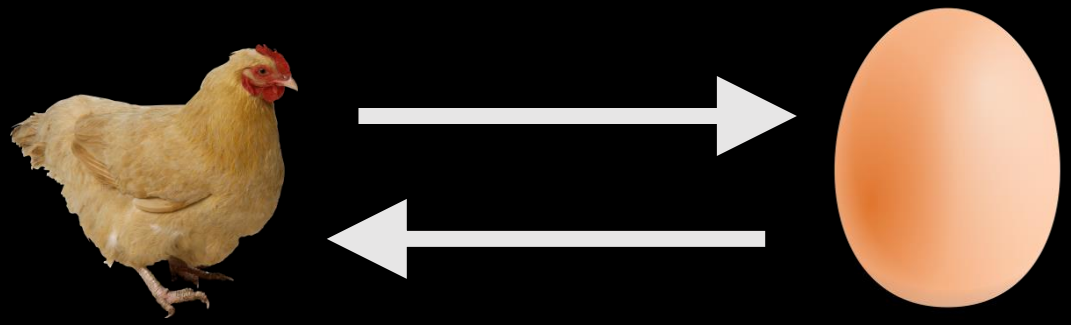


Non-normality of errors

GEVAAR 2

SIMULTANEÏTEIT (WEDERKERIGHEID)

GEVAAR 2: SIMULTANEÏTEIT



Tweerichtingseffect of wederkerig effect

GEVAAR 2: SIMULTANEÏTEIT

Wederkerige effecten of tweerichtings-invlo
(*mutual influence*)

Gevaar

Vertekende schatting van coëfficiënten (en dus verkeerde inschatting van de omvang van het effect)

Opsporing

Goed nadenken. Er is geen (statistische) test die je kan beschermen tegen simultaneïteit

Oplossing

Tijdseries-analyse
Instrumentele variabelen

} Niet in dit vak

8 GEVAREN VAN REGRESSIE

DANGER!!

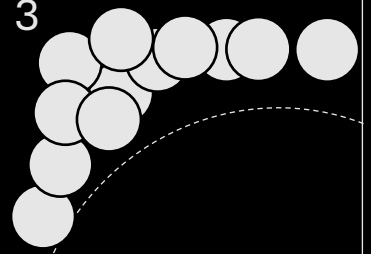
1

Schijnverband

2

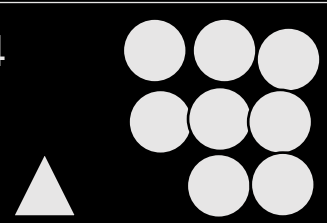
Wederkerigheid / Simultaniteit

3



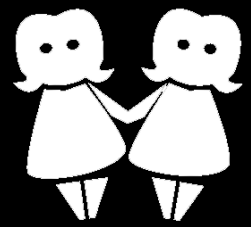
Non-Lineairiteit

4



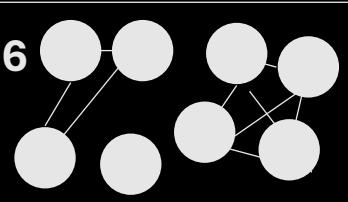
Extreme waarden

5



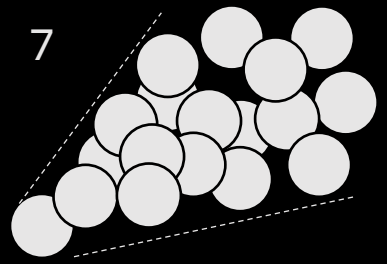
Multicollineariteit

6



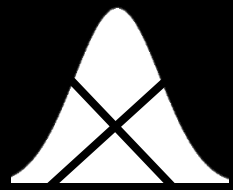
Niet onafhankelijke residuen

7



Heteroskedasticiteit

8



Non-normality of errors

https://elmarjansen.nl/os

OEFENING 1

Het effect van gemiddelde grootte van huishoudens in een buurt
op gemiddeld aantal auto’s per huishouden in die buurt

8 GEVAREN VAN REGRESSIE

DANGER!!

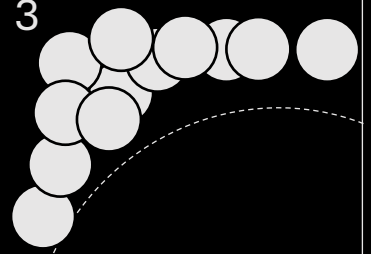
1

Schijnverband

2

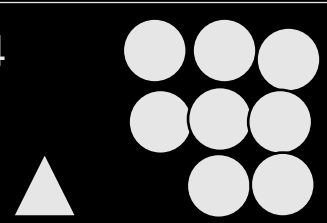
Wederkerigheid / Simultaniteit

3



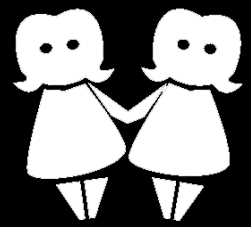
Non-Lineairiteit

4



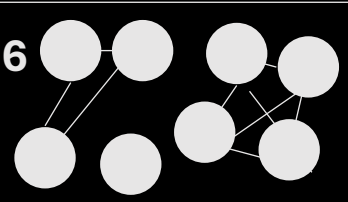
Extreme waarden

5



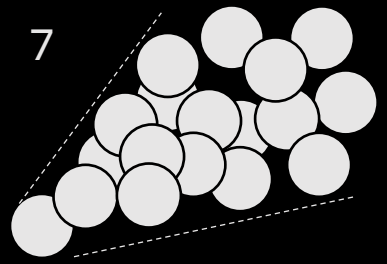
Multicollineariteit

6



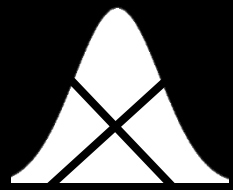
Niet onafhankelijke residuen

7

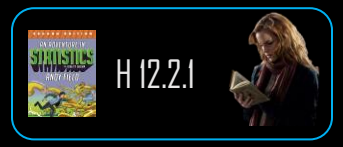


Heteroskedasticiteit

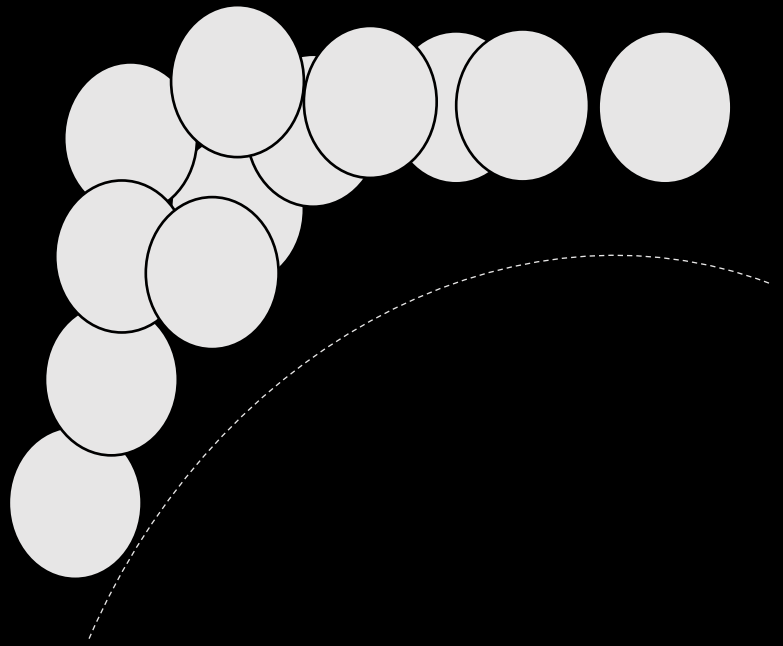
8



Non-normality of errors



GEVAAR 3

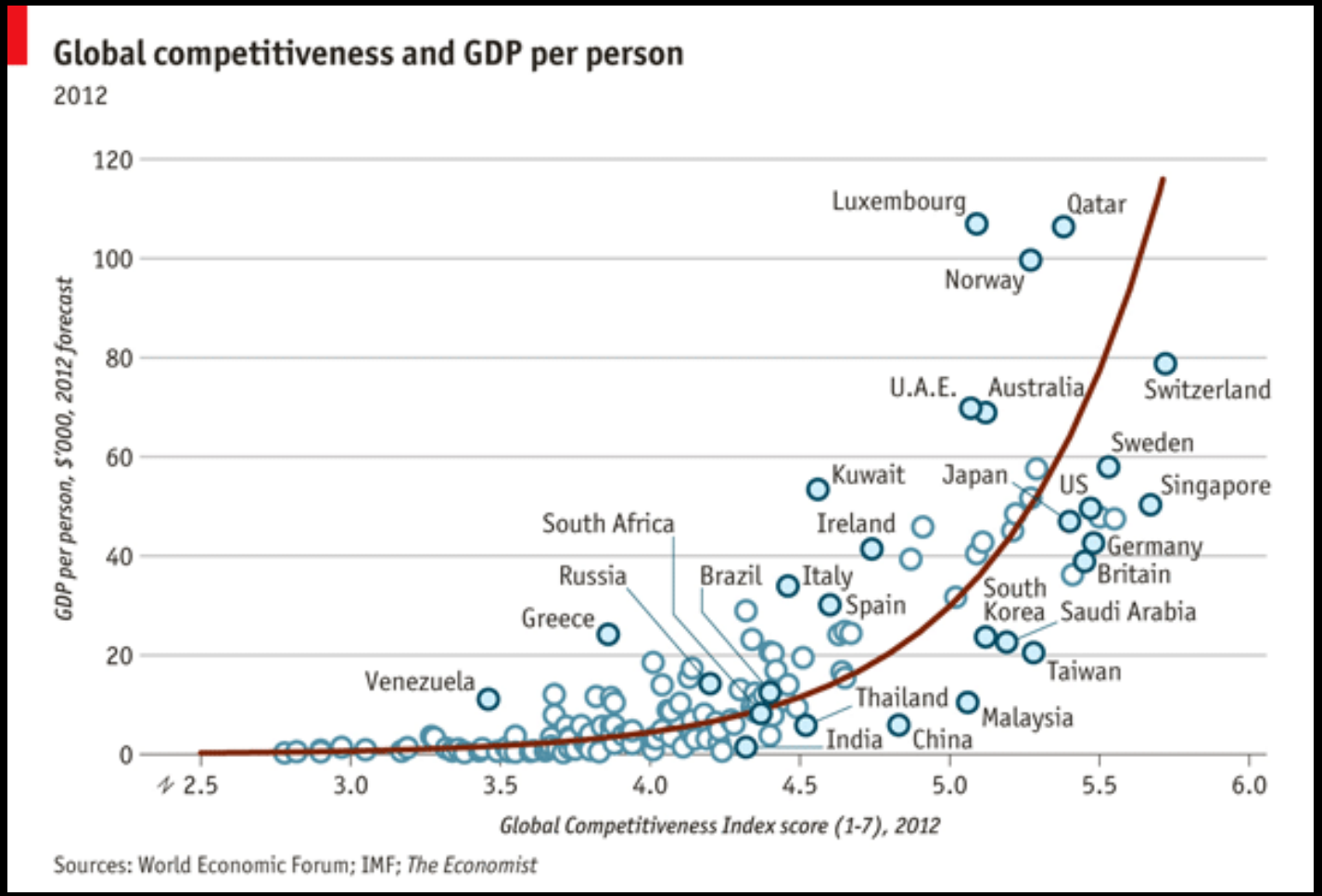


NON-LINEAIRITEIT

GEVAAR 3 NON-LINEAIRITEIT

Het uitgangspunt van lineaire regressie is:

- dat relaties (bij benadering) lineair zijn
- dat de variabelen (bij benadering) continu en interval zijn



GEVAAR 3 NON-LINEAIRITEIT

Je denkt dat er (bijna) geen verband is,
terwijl het er wel is – alleen niet lineair.

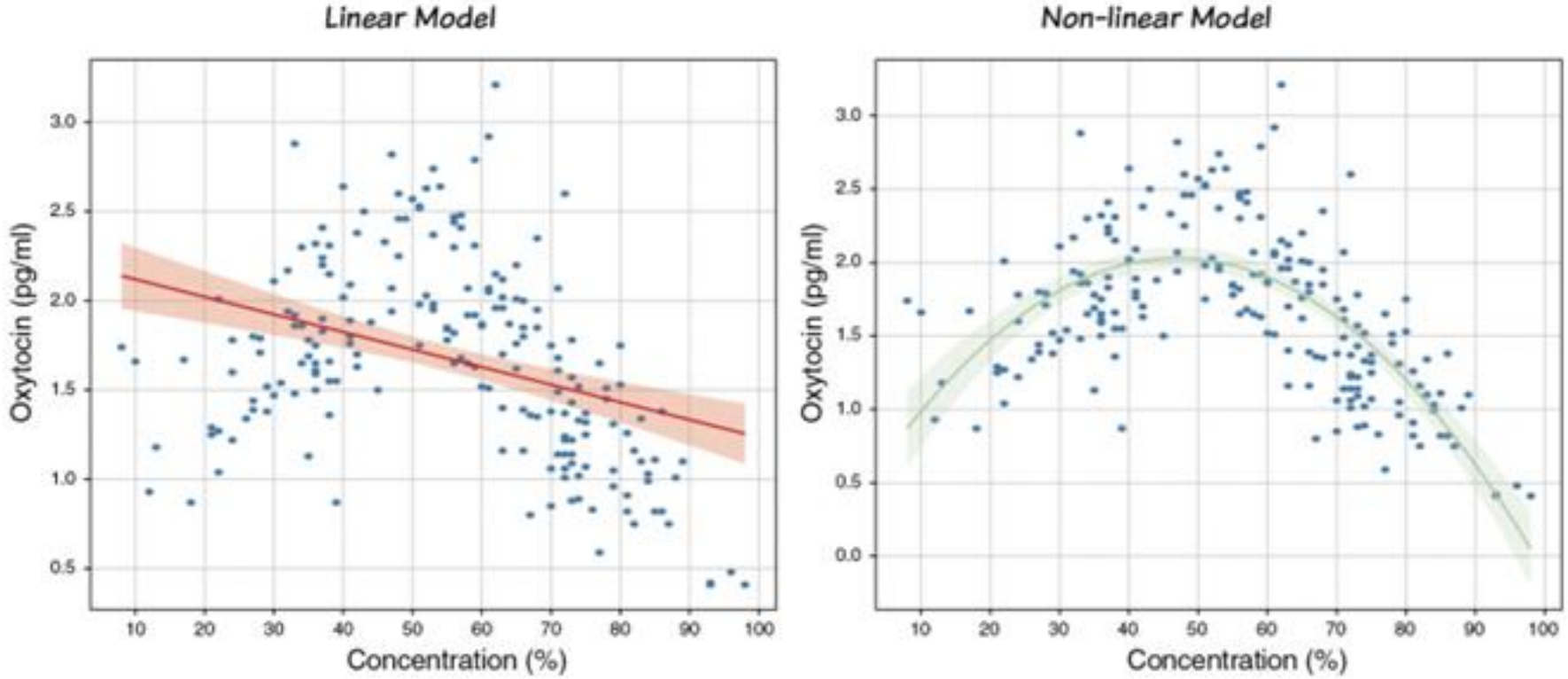
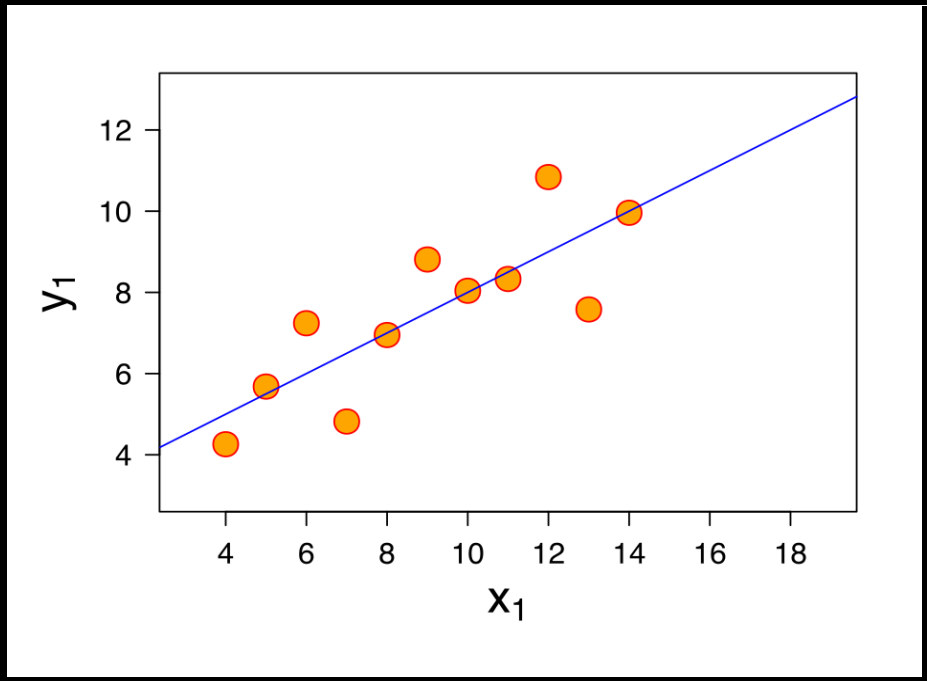


Figure 12.3 Two models fitted to the same data

ALLEMAAL DEZELFDE REGRESSIELIJN…

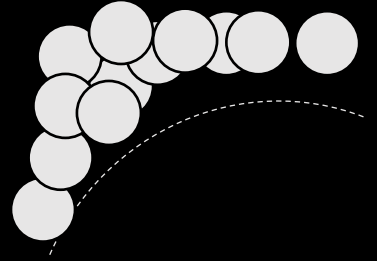


TRANSFORMEREN

Transformatie kan een oplossing zijn:

- We vervangen de variabele door dezelfde variabele waarop een wiskundige bewerking (transformatie) is uitgevoerd
- Veel voorkomende transformaties:
  - **Logaritme** (vooral bij ratio-variabelen als inkomen of aantal inwoners)
  - **Kwadraat** (vooral bij parabolisch effect: dan kwadraat onafhankelijke variabele *en* ongetransformeerde variabele in model)
  - **Wortel**
  - **Relatieve waarde** (waarde gedeeld door een totaal)

GEVAAR 3 NON-LINEAIRITEIT



Aanname van (lineair!) regressie-model is dat alle effecten lineair zijn

Gevaar

geen of alleen kleine effecten vinden, terwijl er wel een relatie is (alleen niet lineair)

Opsporing

Spreidingsdiagram (scatterplot)

Bij meervoudige regressie: spreidingsdiagram tussen residuen en onafhankelijke variabelen

Oplossing

Transformeer variabelen (kwadraat, logaritme etc.)

Ander type regression voor specifieke verdeling afhankelijke variabele (bijv. Poisson of negative binomial; niet in dit vak besproken)

8 GEVAREN VAN REGRESSIE

DANGER!!

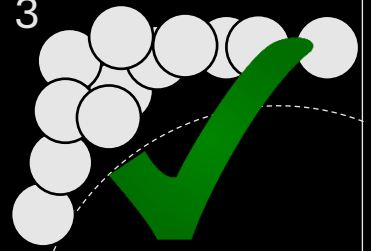
1

Schijnverband

2

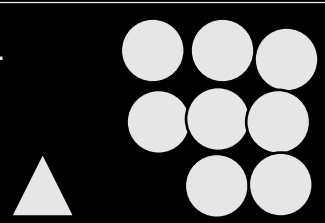
Wederkerigheid / Simultaniteit

3



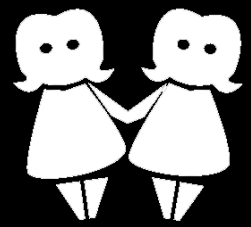
Non-Lineairiteit

4



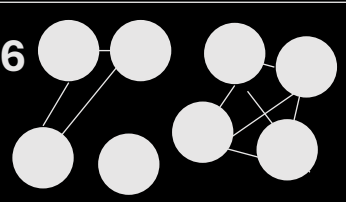
Extreme waarden

5



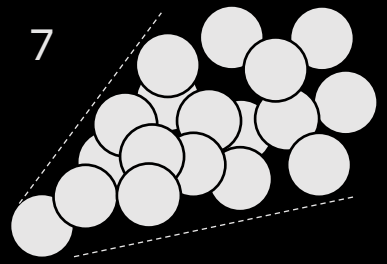
Multicollineariteit

6



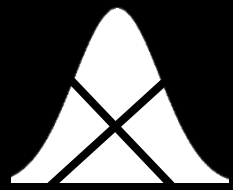
Niet onafhankelijke residuen

7



Heteroskedasticiteit

8



Non-normality of errors

https://elmarjansen.nl/os

OEFENING 2

Het effect van gemiddelde grootte van huishoudens in een buurt
op gemiddeld aantal auto's per huishouden in die buurt

8 GEVAREN VAN REGRESSIE

DANGER!!

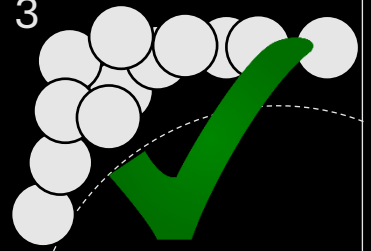
1

Schijnverband

2

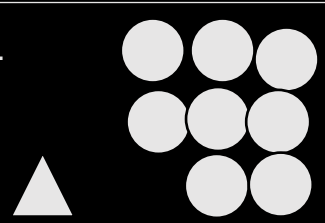
Wederkerigheid / Simultaniteit

3



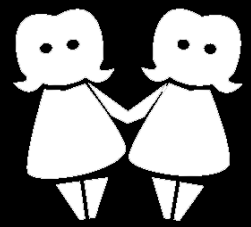
Non-Lineairiteit

4



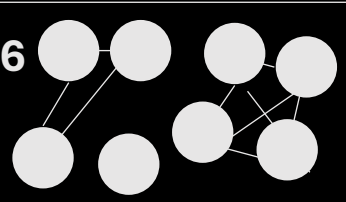
Extreme waarden

5



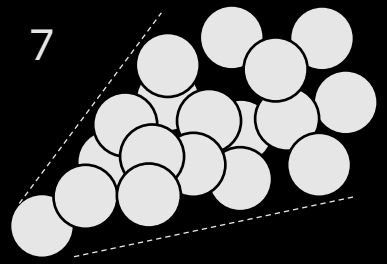
Multicollineariteit

6



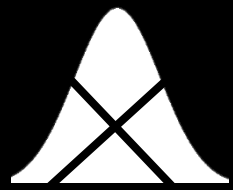
Niet onafhankelijke residuen

7



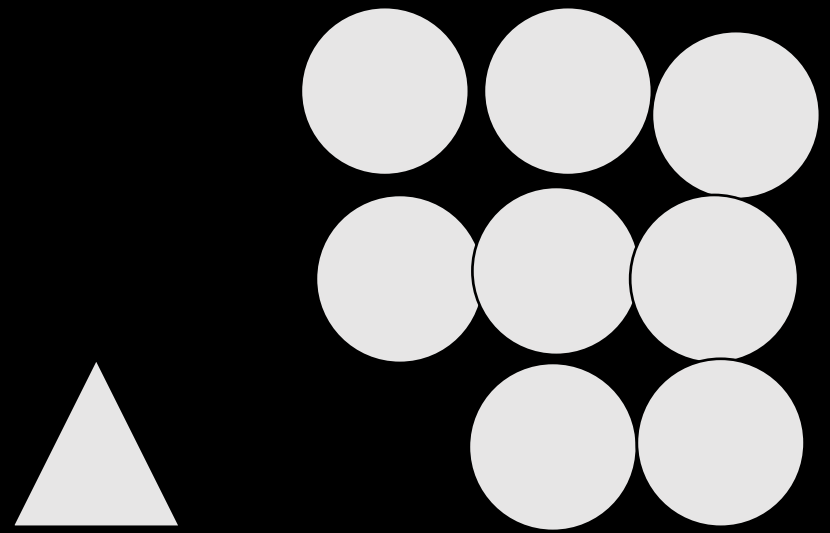
Heteroskedasticiteit

8



Non-normality of errors

GEVAAR 4



EXTREME WAARDEN

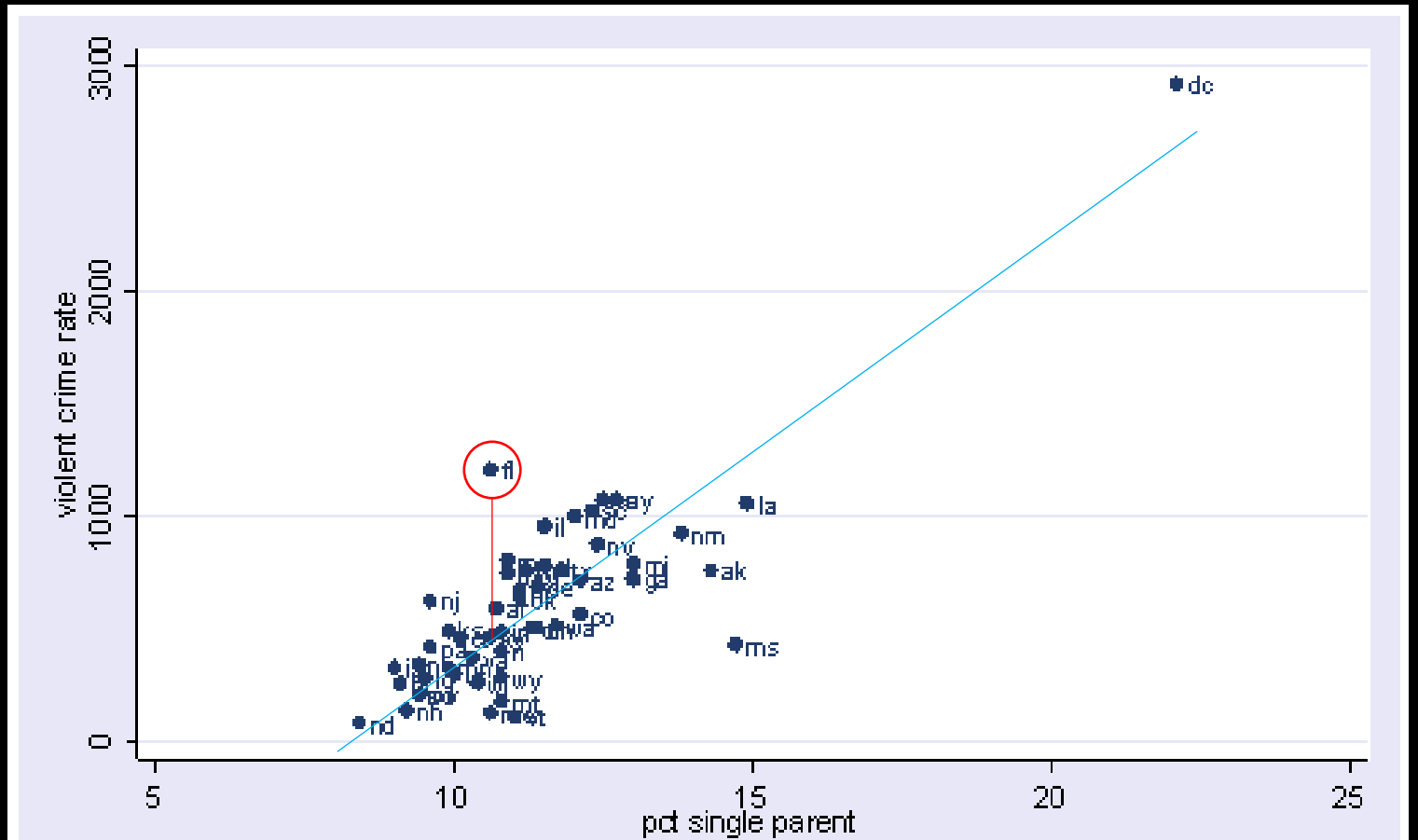
EXTREME WAARDEN

- Regressie kan zeer gevoelig zijn voor **extreme waarden**
- Een **enkele observatie** met zeer **uitzonderlijke waarden** kan **heel veel invloed** hebben op het gevonden effect
- Dit gevaar is het grootst wanneer de observatie veel “**leverage**” (hefboom-werking) heeft

GEVAAR 4
EXTREME WAARDEN

Outlier x Leverage = Influence

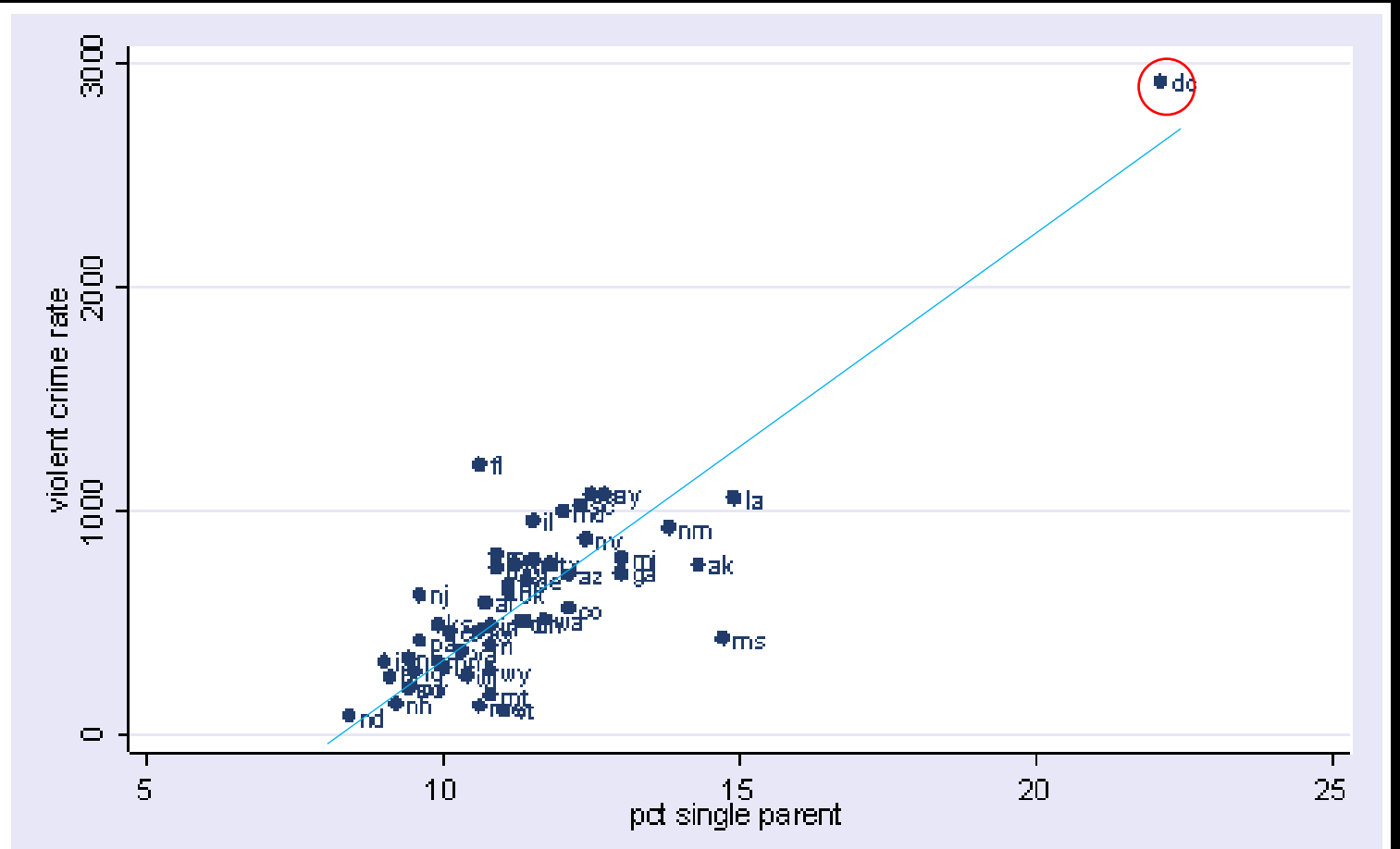
Observatie met groot residu –
dus ver van de regressielijn



GEVAAR 4
EXTREME WAARDEN

Outlier x Leverage = Influence

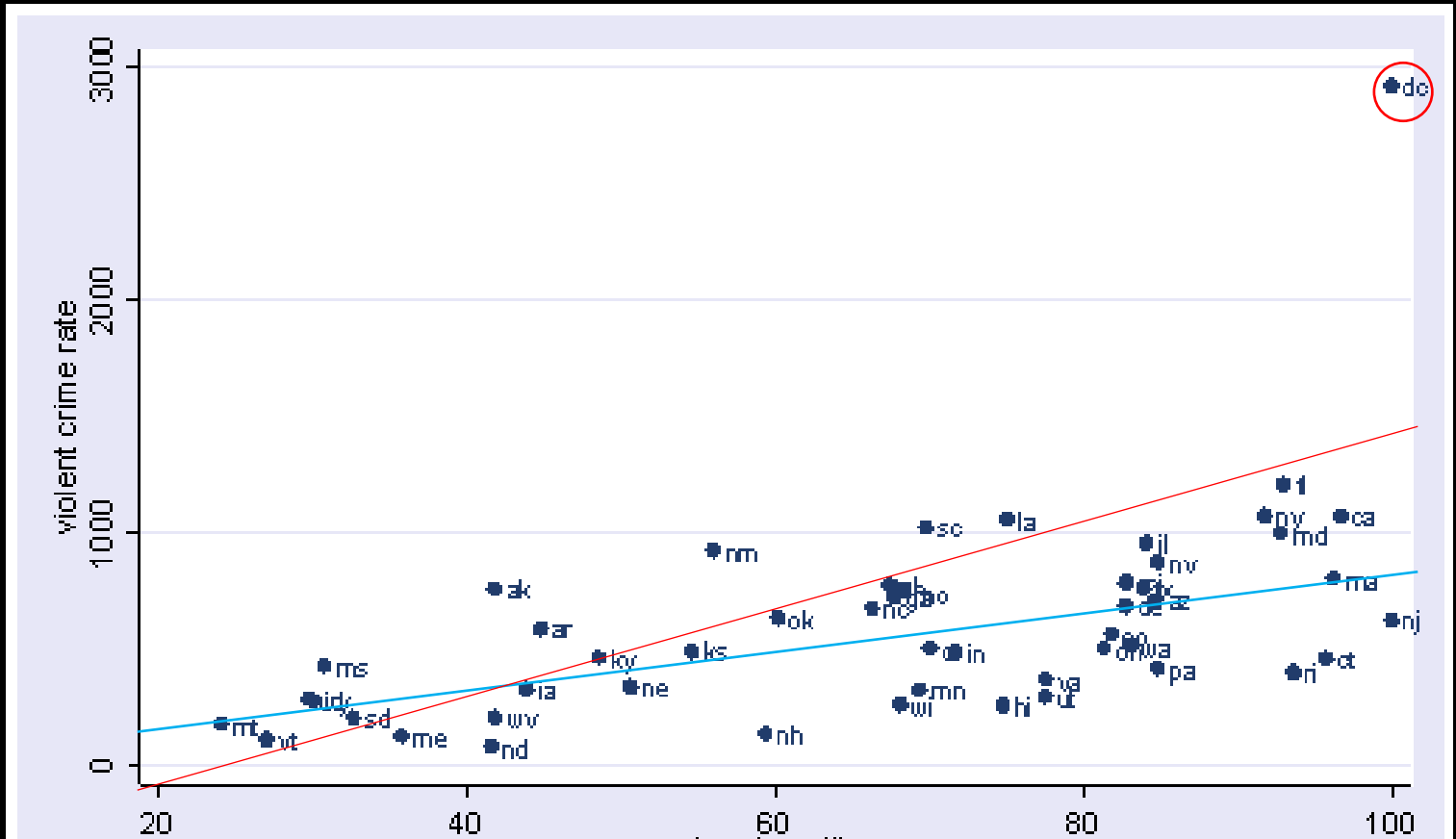
Observatie met extreme waarde op x



GEVAAR 4
EXTREME WAARDEN

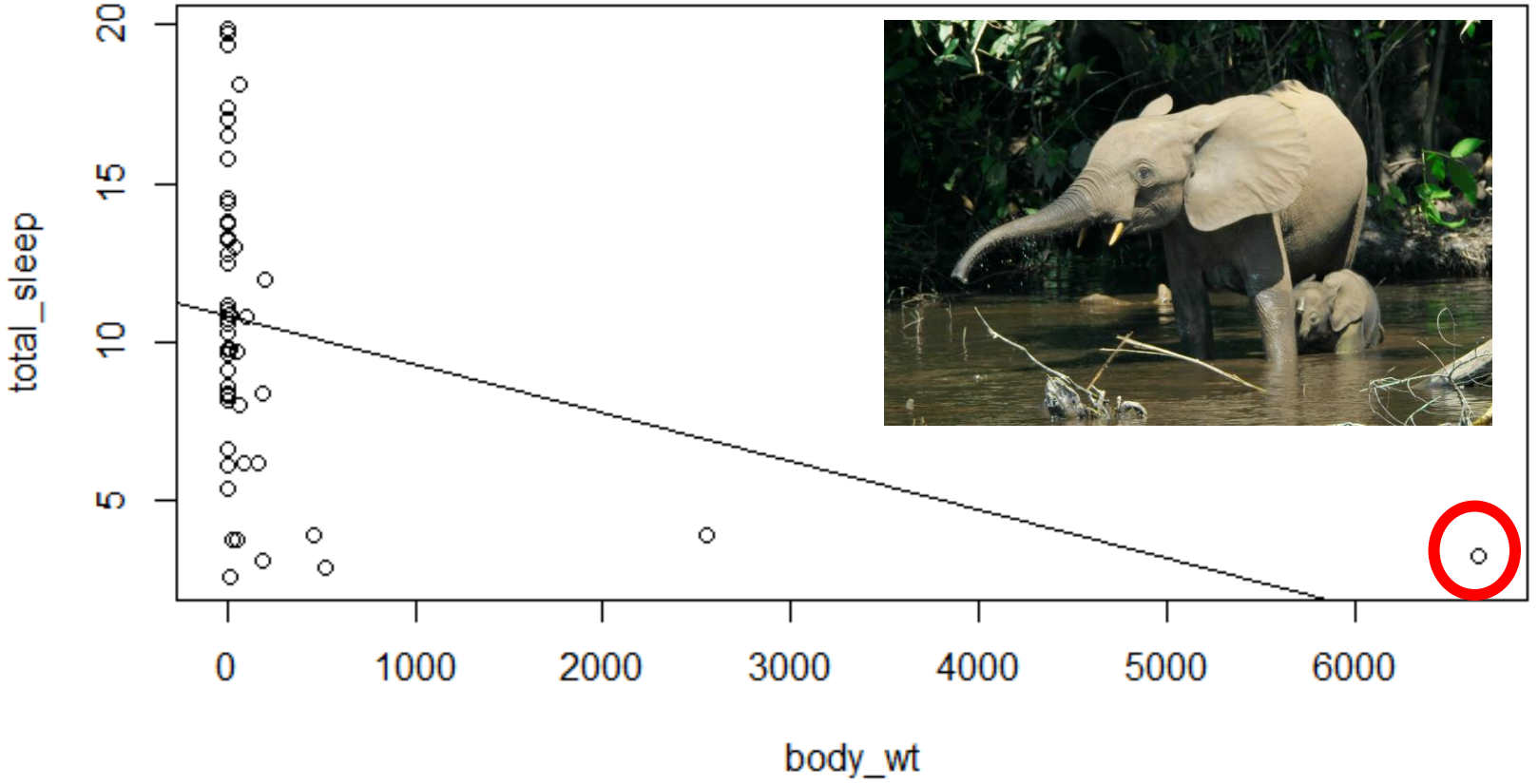
Outlier x Leverage = Influence

Heeft beide eigenschappen
en dus veel invloed op resultaat



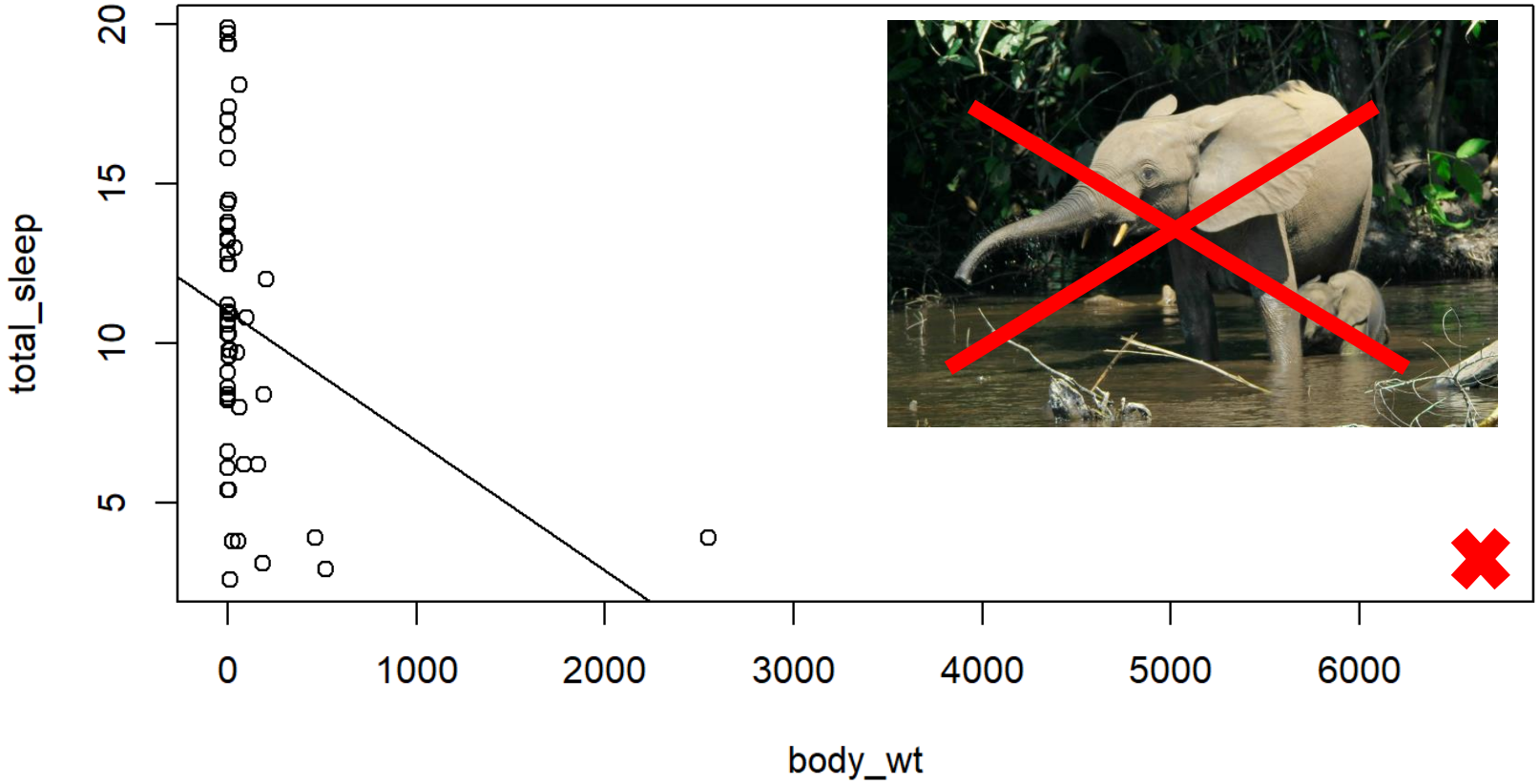
GEVAAR 4
EXTREME WAARDEN

$$slaap_i = 10{,}8 - 0{,}0015 gewicht_i + \varepsilon_i$$

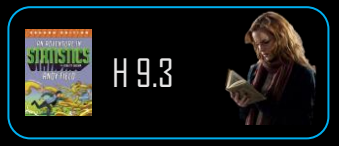


GEVAAR 4
EXTREME WAARDEN

$$slaap_i = 11{,}0 - 0{,}0041 gewicht_i + \varepsilon_i$$



GEVAAR 4
EXTREME WAARDEN



Gevaar

resultaten worden bepaald door maar één of enkele extreme waarnemingen

Opsporing

bestuderen spreidingsdiagram en z-scores

Oplossing

- Eerst: denk na!
  - Kloppen je metingen?
  - Is er een theoretische reden voor je extreme waarden?
  - Moet je misschien transformeren?
- Gooi nooit *zomaar* waarneming weg omdat het een outlier is
- Met beleid waarnemingen weggooien:
  - *winsoring* of *trimming*
- Nerdy oplossingen: *robust regression* (`rlm` in `MASS` library) of *bootstrapping*

8 GEVAREN VAN REGRESSIE

DANGER!!

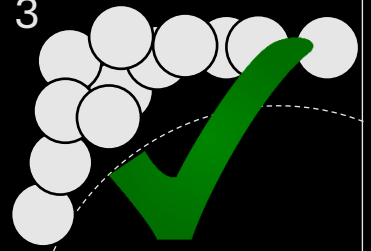
1

Schijnverband

2

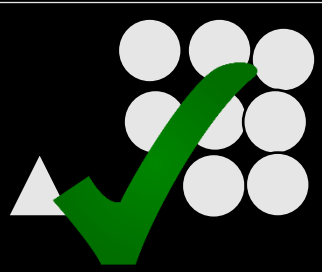
Wederkerigheid / Simultaniteit

3



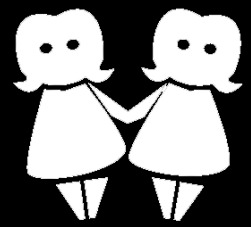
Non-Lineairiteit

4



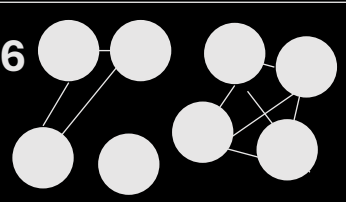
Extreme waarden

5



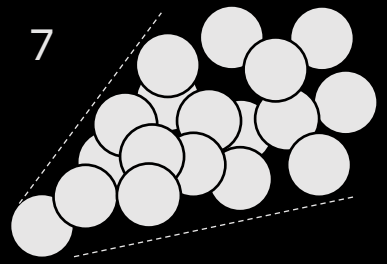
Multicollineariteit

6



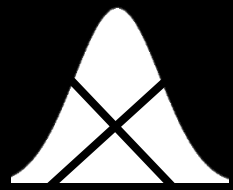
Niet onafhankelijke residuen

7



Heteroskedasticiteit

8

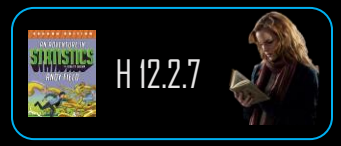


Non-normality of errors

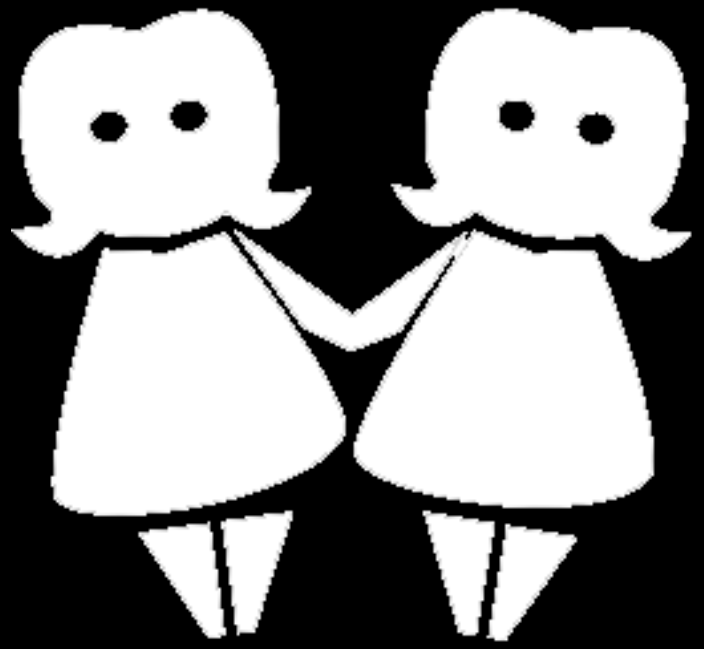
https://elmarjansen.nl/os

OEFENING 3

Het effect van gemiddelde grootte van huishoudens in een buurt
op gemiddeld aantal auto’s per huishouden in die buurt

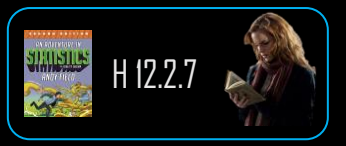


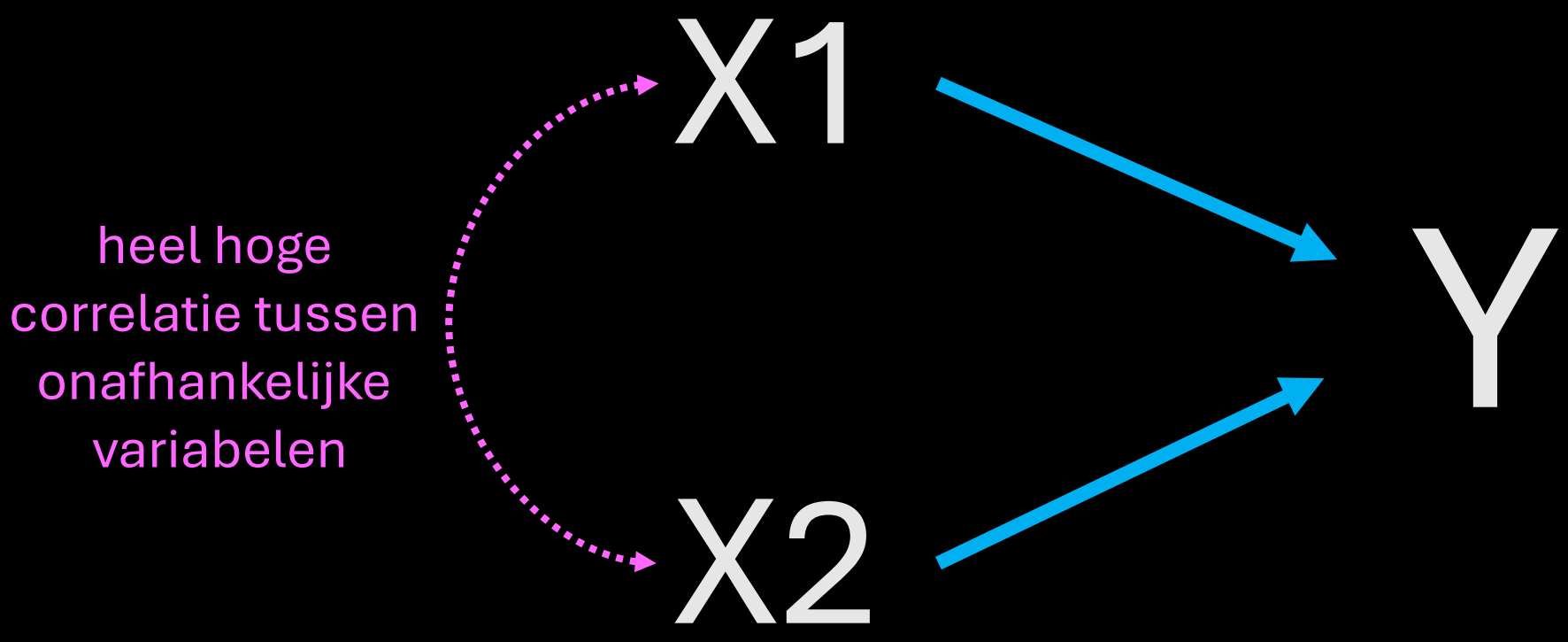
GEVAAR 5



MULTICOLLINEARITEIT

MULTICOLLINEARITEIT

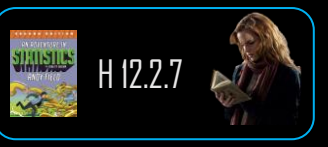


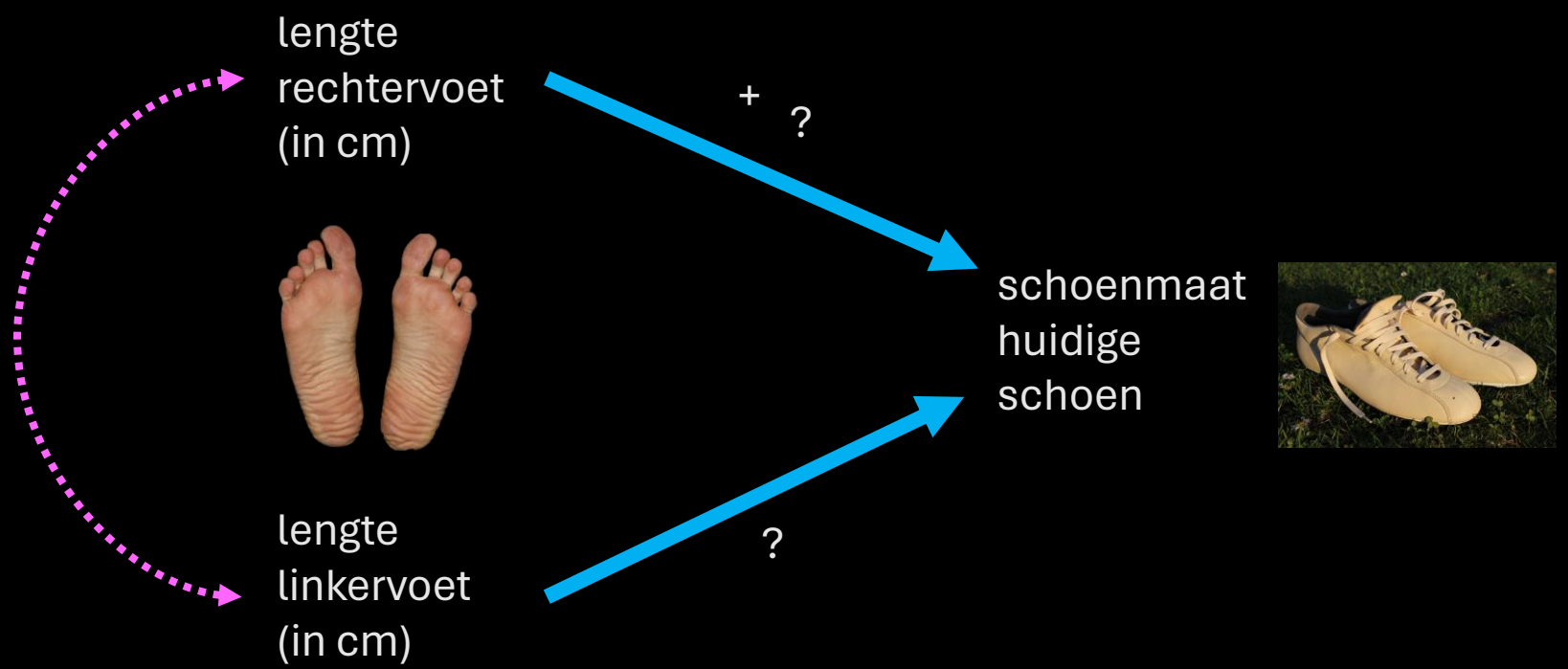


MULTICOLLINEARITEIT: VOORBEELD

$$schoenmaat_i = \beta_0 + \beta_1 rvoet_i + \varepsilon_i$$

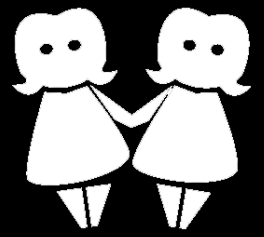
$$schoenmaat_i = \beta_0 + ?\beta_1 rvoet_i + ?\beta_2 lvoet_i + \varepsilon_i$$





Het is nu onmogelijk om nog te bepalen wat de afzonderlijke waarden van β_1 en β_2 zijn – omdat die effecten niet van elkaar te scheiden zijn

GEVAAR 5 MULTICOLLINEARITEIT



Meerdere onafhankelijke variabelen verklaren (bijna) hetzelfde deel van de variatie in de afhankelijke variabele

Gevaar

→ Rare (en moeilijk te interpreteren) coëfficieten
→ Grote (“opgeblazen”of “inflated”) standaardfout (= enorme uncertainty)

Opsporing

bereken *vif* (moet onder 10 zijn)

Variance Inflation Factor

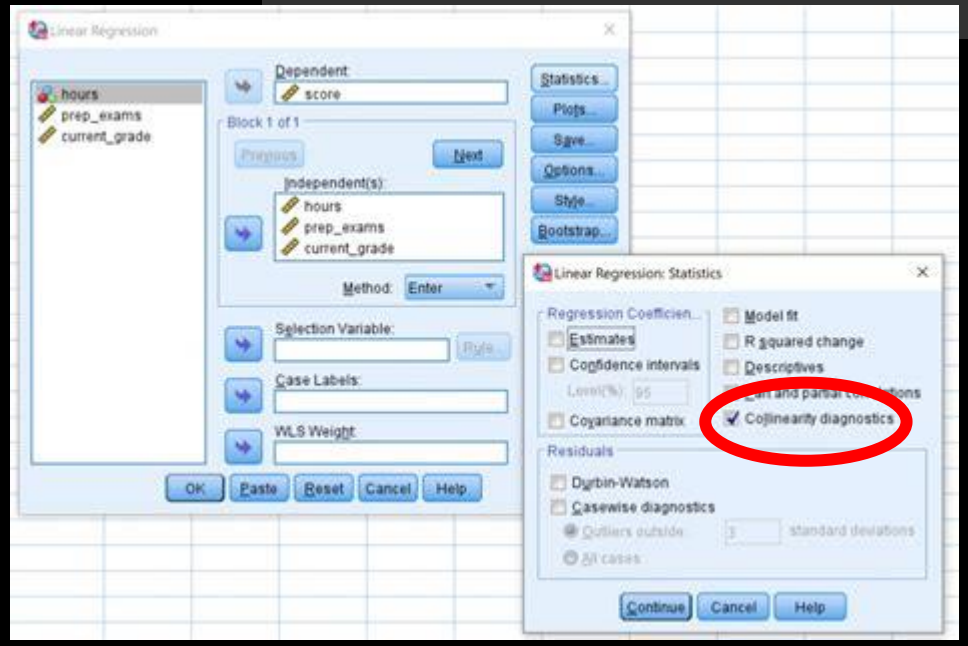
Oplossing

Variabelen uit het model halen
Meerdere variabelen samenvoegen tot één schaal

VIF BEREKENEN IN R EN SPSS

```r
mod <- lm(life_span ~ dreaming + non_dreaming, data=sleep)
library(car)
vif(mod)
```

```
Loading required package: carData
    dreaming non_dreaming
    1.376535     1.376535
```



Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	hours	.856	1.169
	prep_exams	.713	1.403
	current_grade	.657	1.522

a. Dependent Variable: score

8 GEVAREN VAN REGRESSIE

DANGER!!

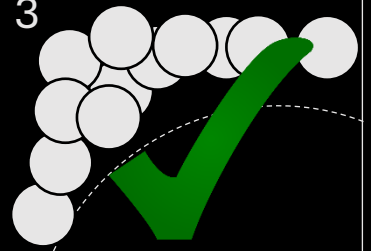
1

Schijnverband

2

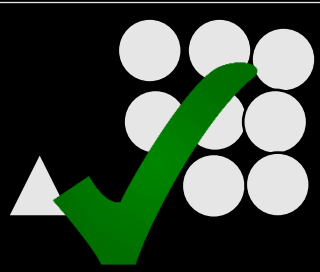
Wederkerigheid / Simultaniteit

3



Non-Lineairiteit

4

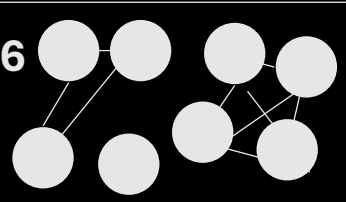


Extreme waarden

5

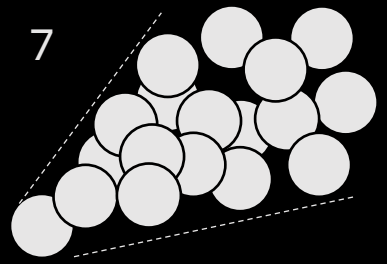
Multicollineariteit

6



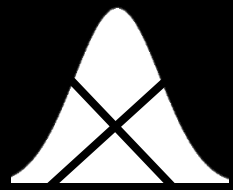
Niet onafhankelijke residuen

7



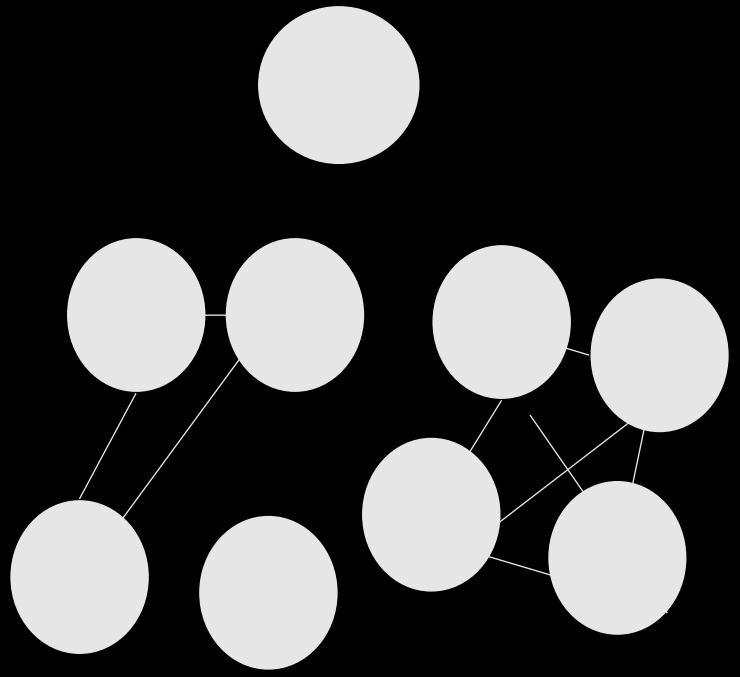
Heteroskedasticiteit

8



Non-normality of errors

GEVAAR 7



NIET-ONAFHANKELIJKE RESIDUEN

GEVAAR 6
NIET-ONAFHANKELIJKE RESIDUEN

De gegevens moeten uit een echte aselecte steekproef komen

Alle observaties moeten dus *onafhankelijk* van elkaar zijn

Gevaar

Onjuiste (te lage) standaard-fouten: onderschatting van onzekerheid

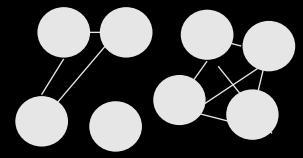
Opsporen van problemen

Nadenken! Geen statistische manier om achter te komen

Voorbeelden
- Tijdserie
- Leerlingen van verschillende scholen

Oplossingen? Niet met normale (OLS) -regressie

Andere methoden:
- Tijd-series
- Multilevel-analyse
- Paired samples T-test
- Dummies voor groepen (“Fixed effects model”)



8 GEVAREN VAN REGRESSIE

DANGER!!

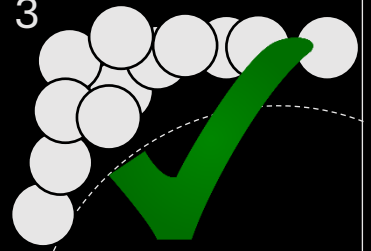
1

Schijnverband

2

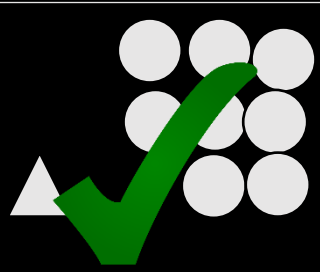
Wederkerigheid / Simultaniteit

3



Non-Lineairiteit

4

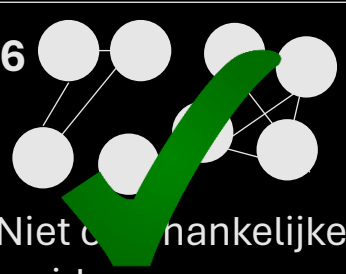


Extreme waarden

5

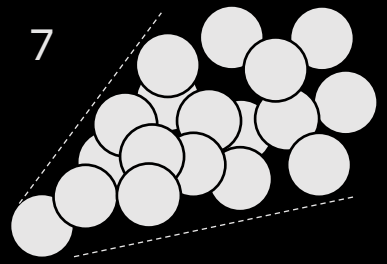
Multicollineariteit

6



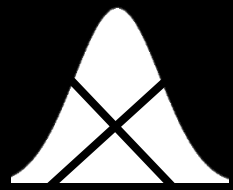
Niet [illegible]nkelijke residuen

7

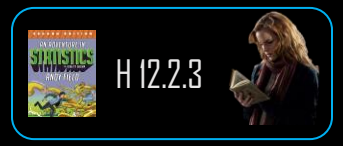


Heteroskedasticiteit

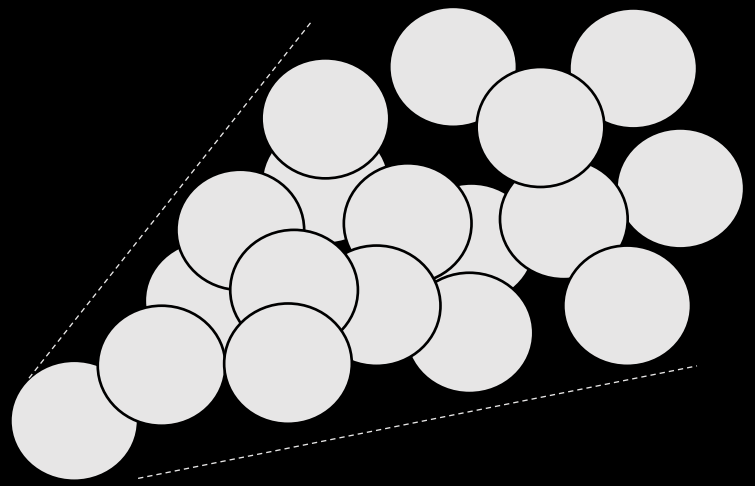
8



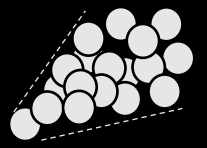
Non-normality of errors



GEVAAR 7



HETEROSKEDASTICITEIT



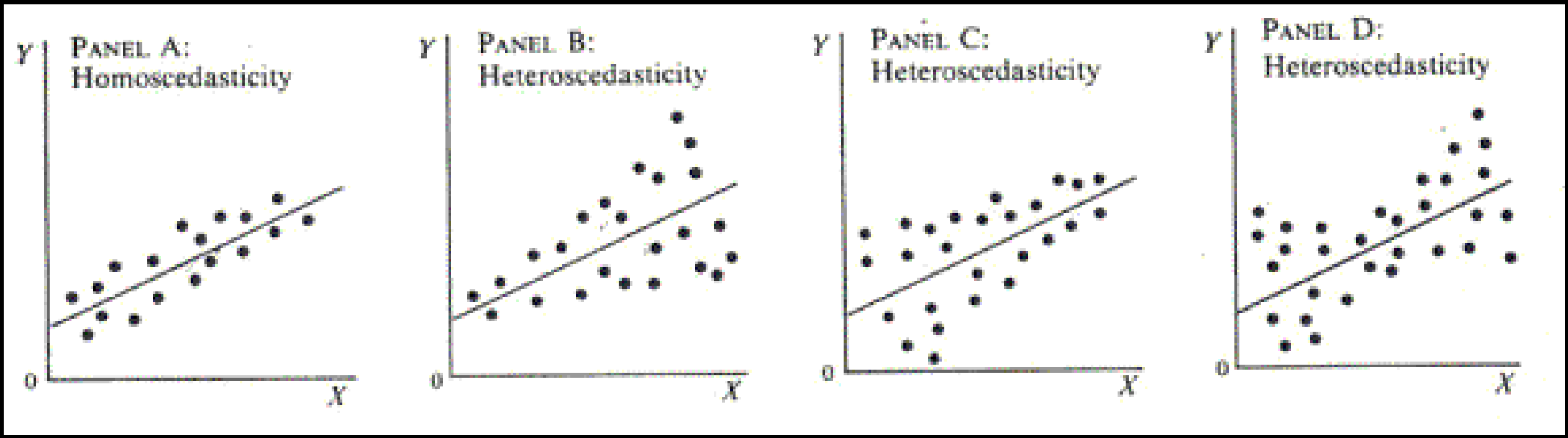
Endogeen

Exogeen

Heteroskedastisch

Homoskedastisch

GEVAAR 7 HETEROSKEDASTICITEIT



H 12.2.3

GEVAAR 7 HETEROSKEDASTICITEIT

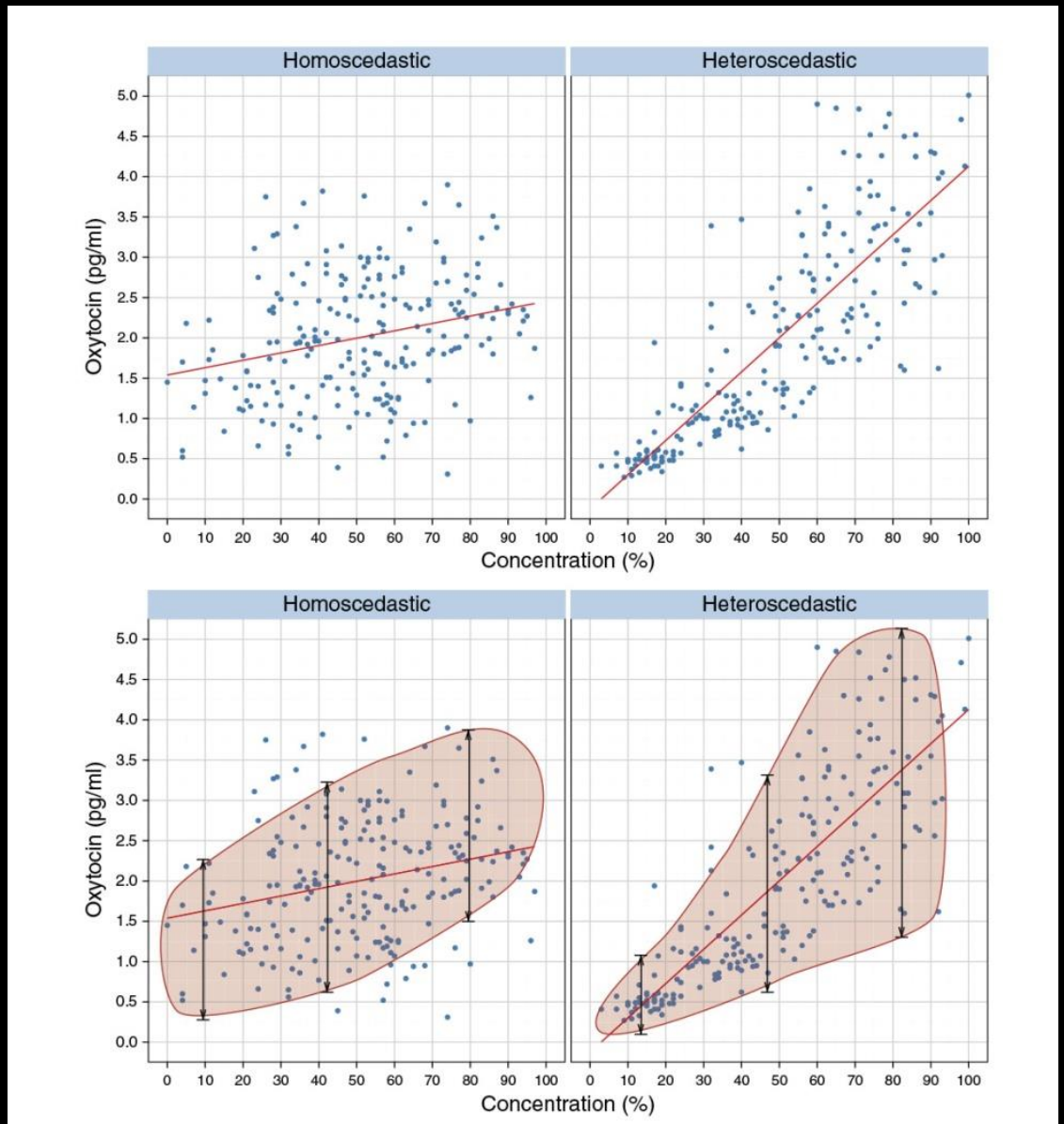
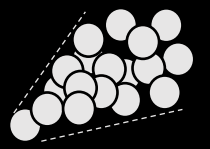
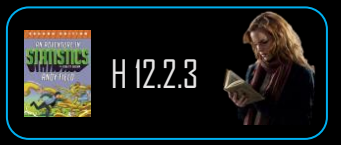


Figure 12.5 *Graphs illustrating data with homogeneous (left) and heterogeneous (right) variances*

GEVAAR 7 HETEROSKEDASTICITEIT





Variantie van residu ϵ_i moet onafhankelijk zijn van de waarde van de onafhankelijke variabelen

Gevaar

onjuiste standaardfout (en dus onjuist betrouwbaarheidsinterval en p-waarde)

Opsporing

spreidingsdiagram met voorspelde waarden ($\hat{y}_i$) en residuen (ϵ_i)

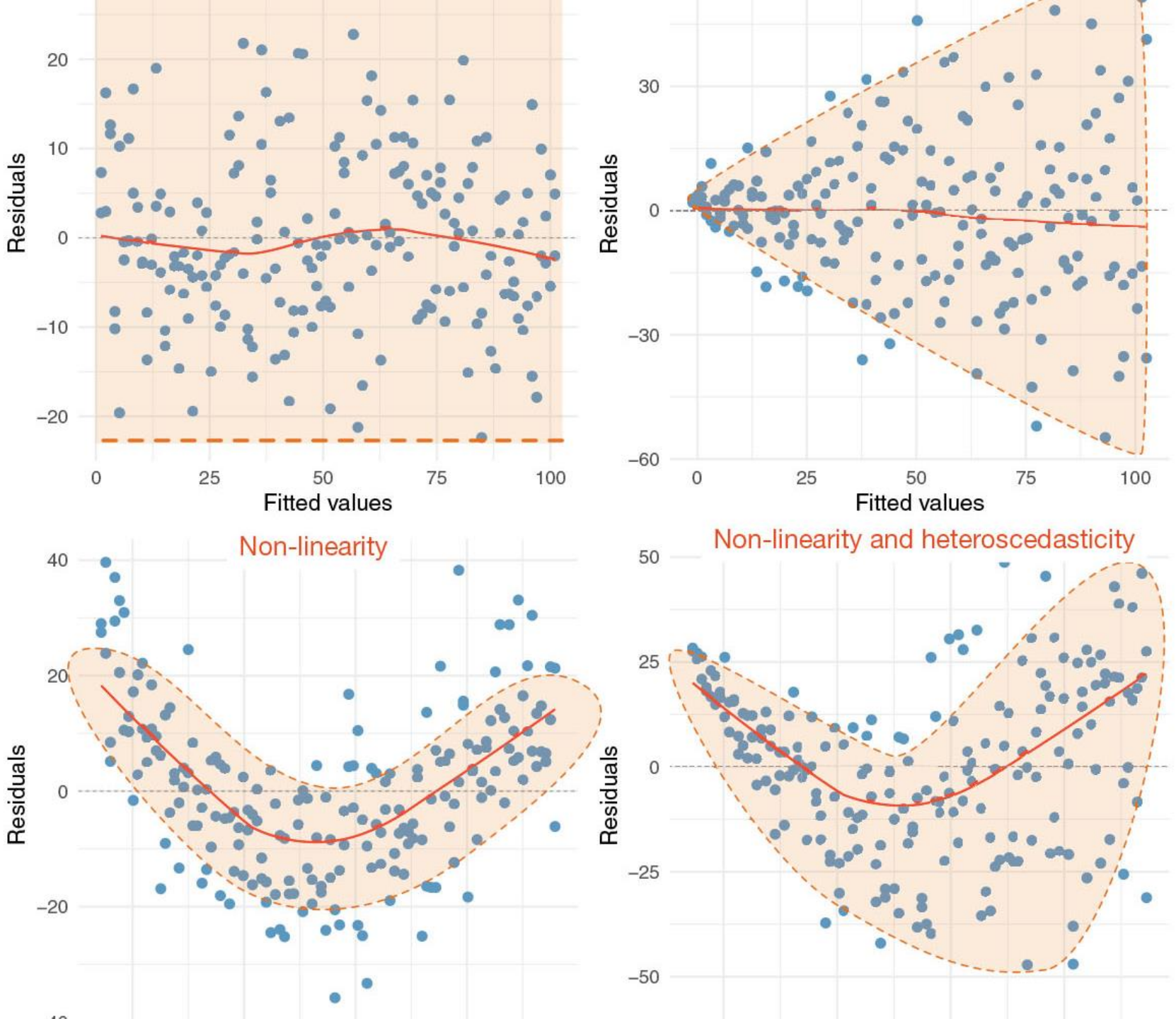
Oplossing

Transformeer afhankelijke variabele (log, kwadraat etc.)

Controlevariabelen toevoegen

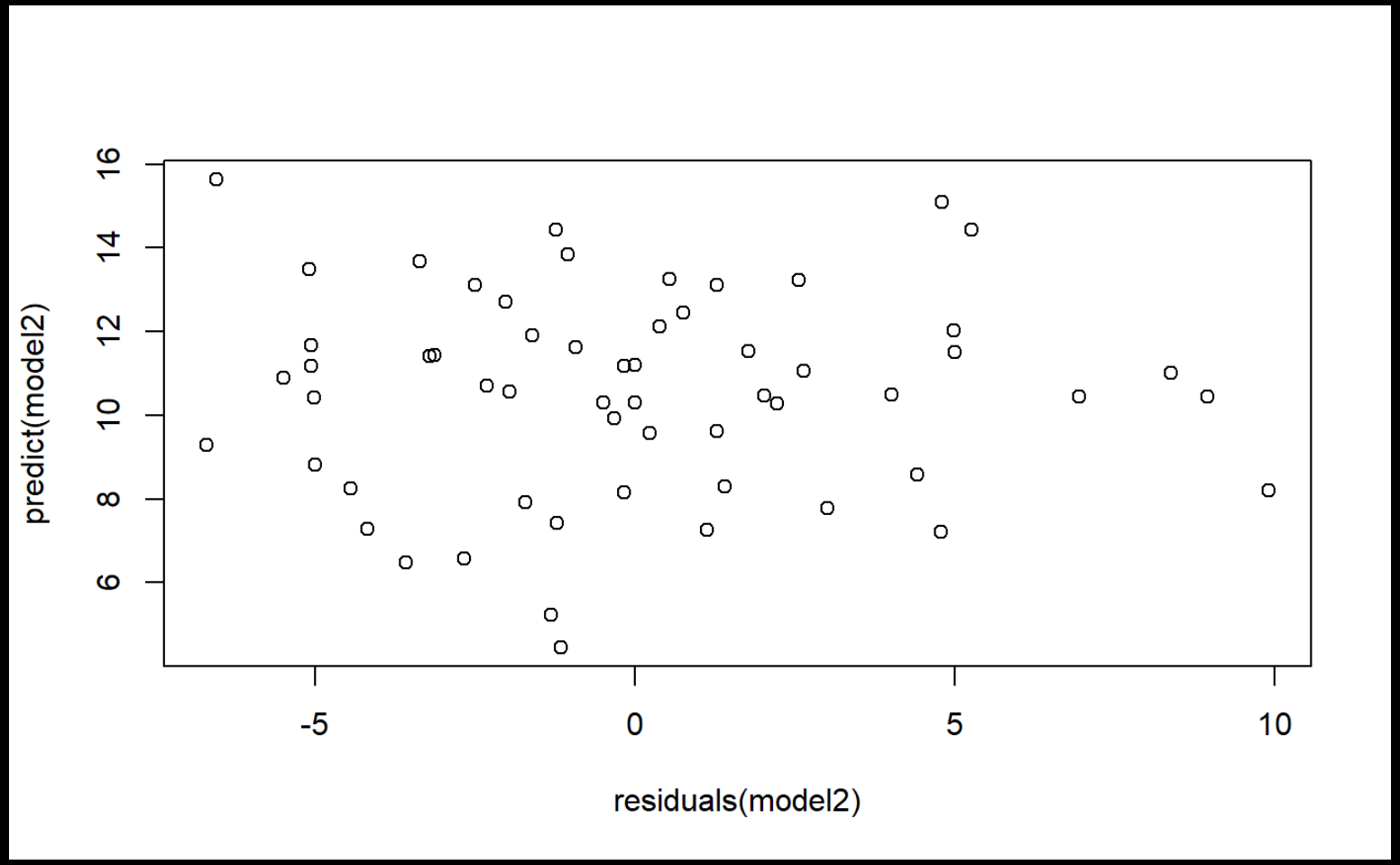
GEVAAR 3, 4 EN 7 OPSPOREN MET RESIDU-PLOT

Zet *voorspelde waarde* $\hat{y}_i$ af tegen *residu* ϵ_i in spreidingdiagram (scatterplot):

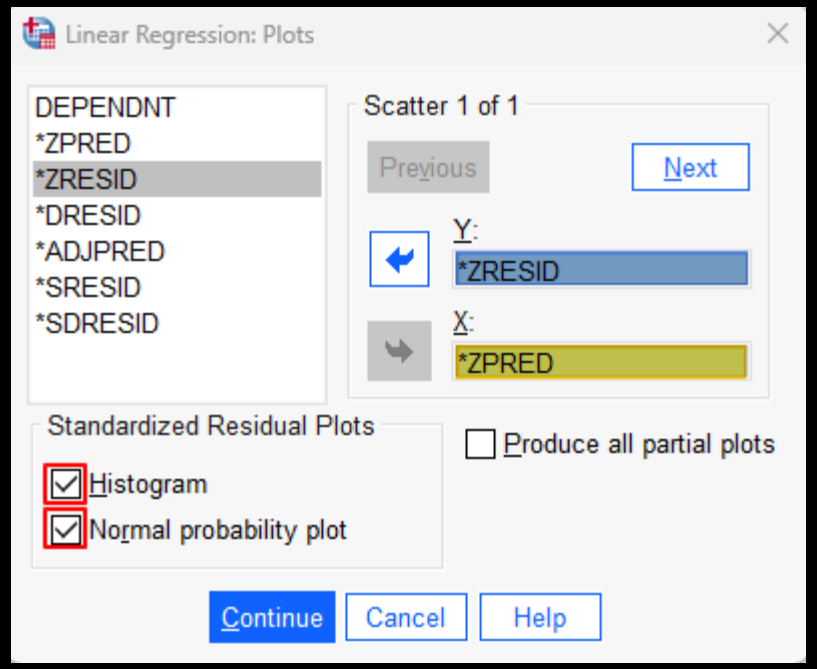

Residuals
Fitted values
Non-linearity
Non-linearity and heteroscedasticity

IN R

```
plot(residuals(model2), predict(model2))
```



IN SPSS


Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
*ZRESID
X:
*ZPRED
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help

8 GEVAREN VAN REGRESSIE

DANGER!!

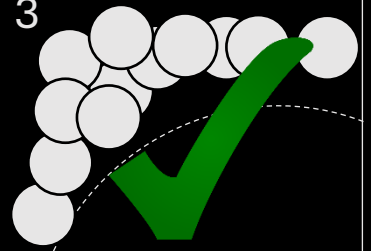
1

Schijnverband

2

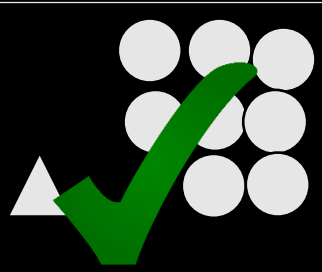
Wederkerigheid / Simultaniteit

3



Non-Lineairiteit

4

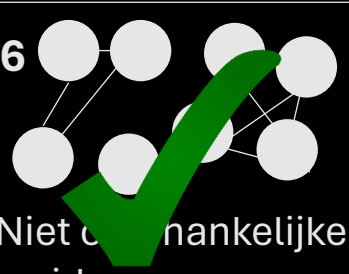


Extreme waarden

5

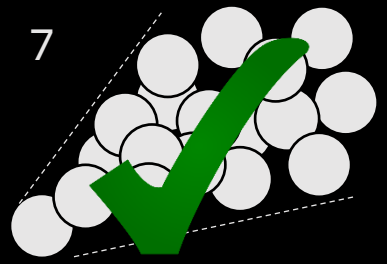
Multicollineariteit

6



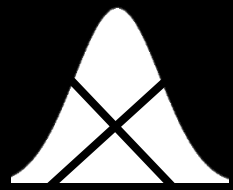
Niet [illegible]nkelijke residuen

7

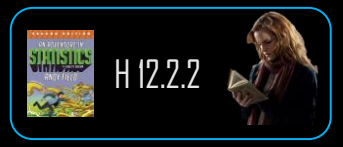


Heteroskedasticiteit

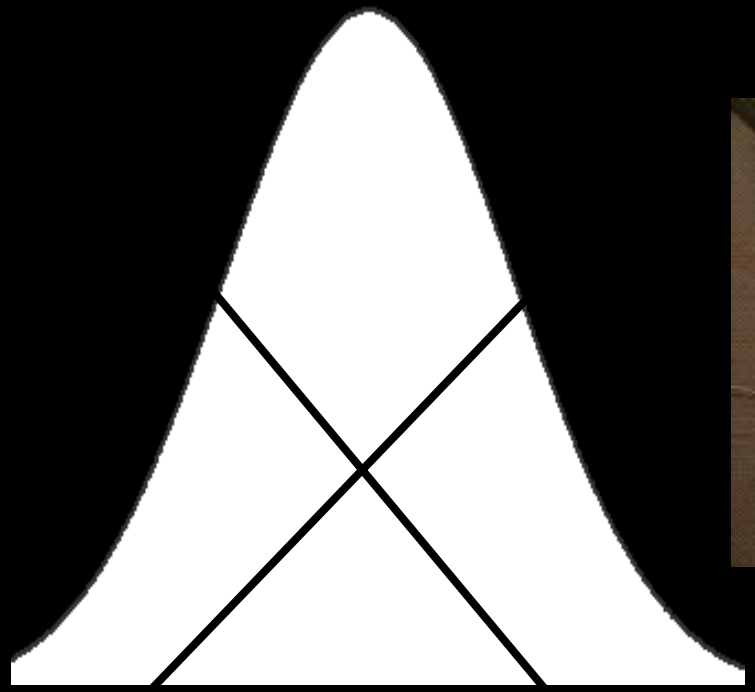
8



Non-normality of errors



GEVAAR 8

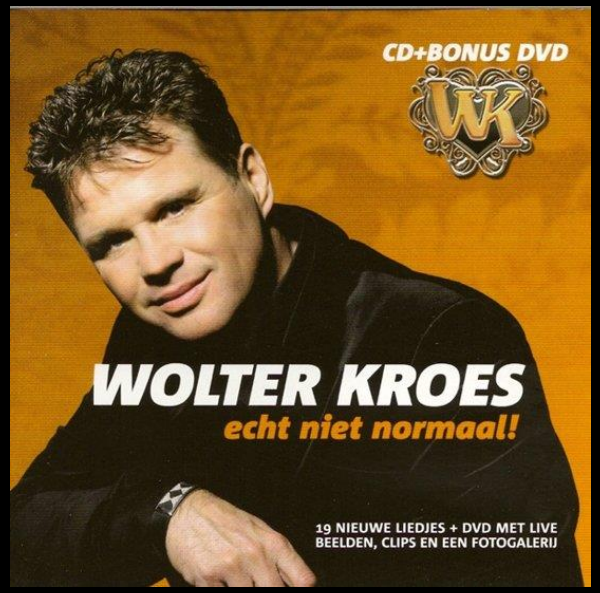


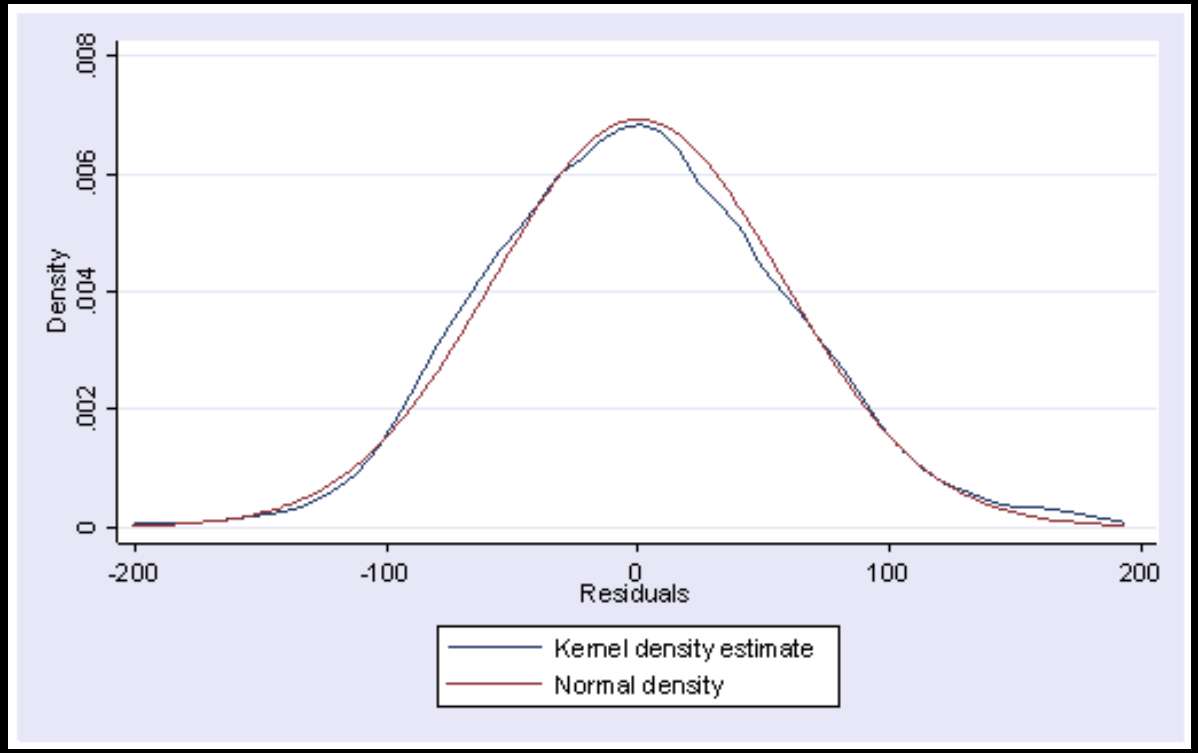
NIET-NORMALITEIT VAN RESIDUEN

GEVAAR 8 NIET-NORMALITEIT

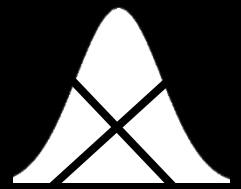
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

residu





GEVAAR 8 NIET-NORMALITEIT



Residuen moeten normaal verdeeld zijn

Gevaar

Onjuiste standaardfouten en significantie, vooral bij kleine steekproef (n < 20)

Opsporing

Bekijk de verdeling van *residuen*.

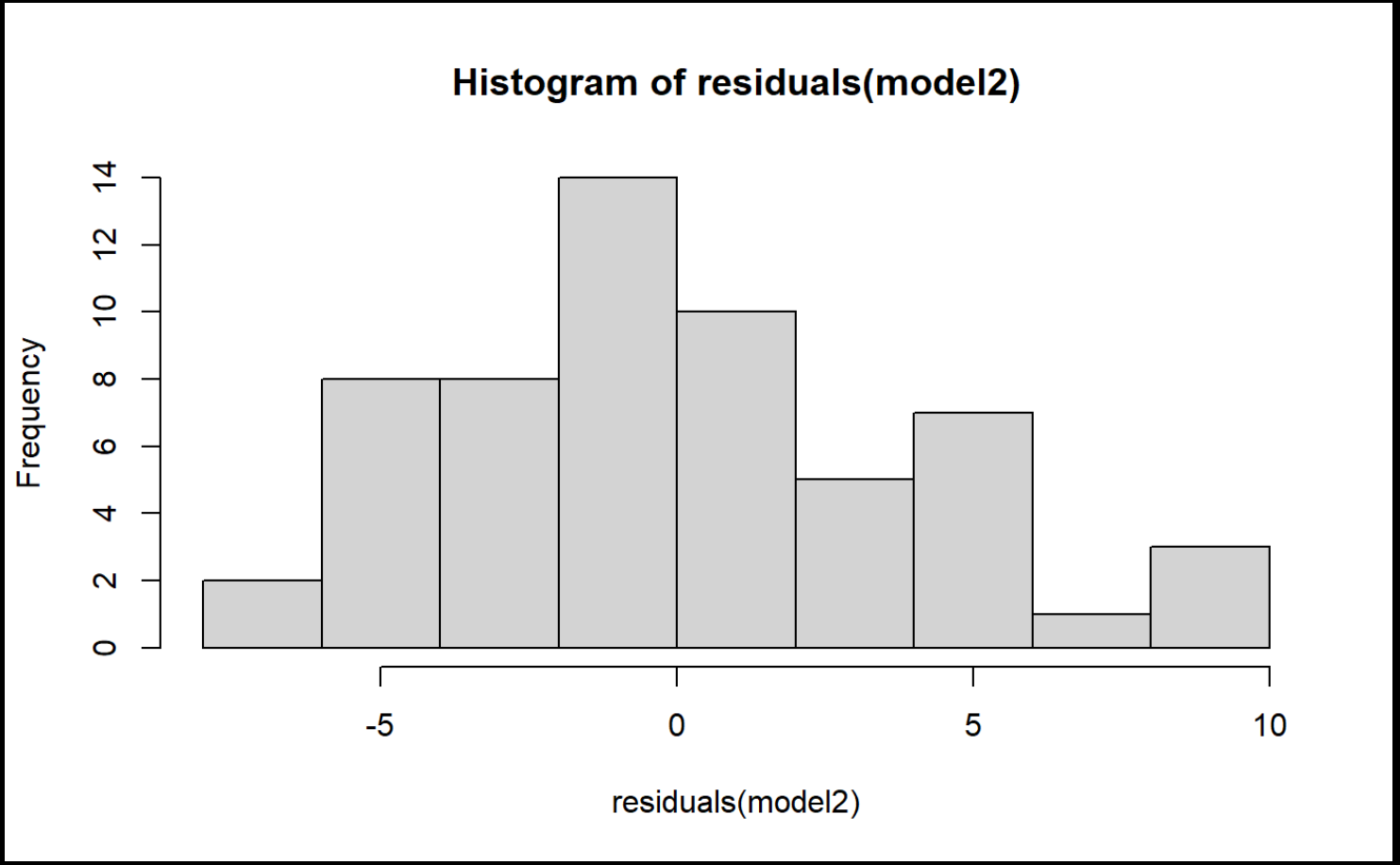
Ziet het er (min of meer) normaal uit?

Oplossing

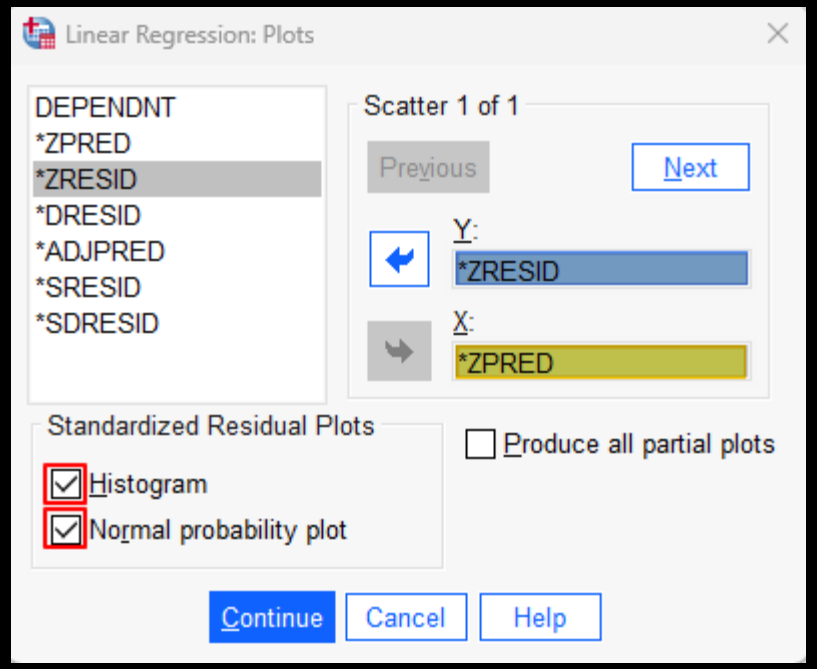
- Transformeer afhankelijke variabele (log, kwadraat, etc.)
- Voeg controle-variabelen toe
- Gebruik “niet parametrische”-tests

HISTOGRAM VAN RESIDUEN IN R

```
hist(residuals(model2))
```



HISTOGRAM VAN RESIDUEN IN SPSS



PSST...

(eigenlijk is dit niet vaak
echt een probleem
– door de Centrale Limietstelling
sowieso alleen bij kleine N -
en dan nog blijkt het probleem
allemaal wel mee te vallen)

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576-2590.

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of clinical epidemiology*, *98*, 146-151.

8 GEVAREN VAN REGRESSIE

DANGER!!

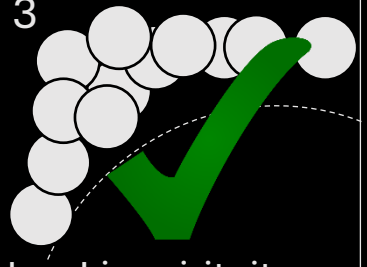
1

Schijnverband

2

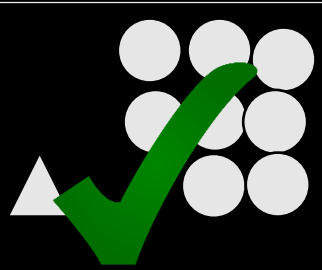
Wederkerigheid / Simultaniteit

3



Non-Lineairiteit

4

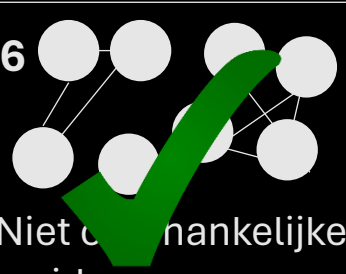


Extreme waarden

5

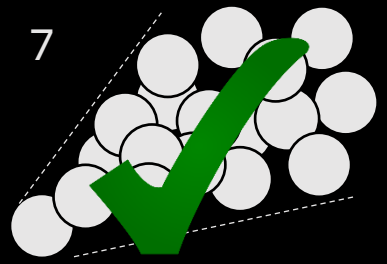
Multicollineariteit

6



Niet [illegible]nankelijke residuen

7



Heteroskedasticiteit

8

Non-Normality of errors

8 DANGERS OF REGRESSION

DANGER!!

1 

Schijnverband

2 

Wederkerigheid / Simultaniteit

3 

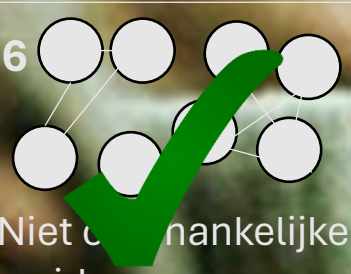
Non-Lineairiteit

4 

Extreme waarden

5 

Multicollinearitеit

6 

Niet [illegible]nkelijke residuen

7 

Heteroskedasticiteit

8 

Non-normality of errors


ALMOST THERE...

8 DANGERS OF REGRESSION

DANGER!!

Endogeniteit

1 Spuriousness

2 Reciprocity

Regression Assumptions

3 Non-Linearity

4 Heteroscedasticity

5 Multicolinearity

6 Non-[illegible]ependent errors

7 Non[illegible]ormality of errors

8 Atypical observations

BLUE

Als dit allemaal goed gaat,
geeft OLS een model dat BLUE is:

Best
Linear
Unbiased
Estimates

Om precies te zijn, geeft OLS volgens de Gauss-Markov stelling BLUE schattingen van de coëfficienten als aan deze voorwaarden is voldaan:

-> Lineairiteit
-> Sphericiteit van residuen
-> Exogeniteit van onafhankelijke variabelen

https://elmarjansen.nl/os

OEFENING 4

Het effect van gemiddelde grootte van huishoudens in een buurt
op gemiddeld aantal auto's per huishouden in die buurt

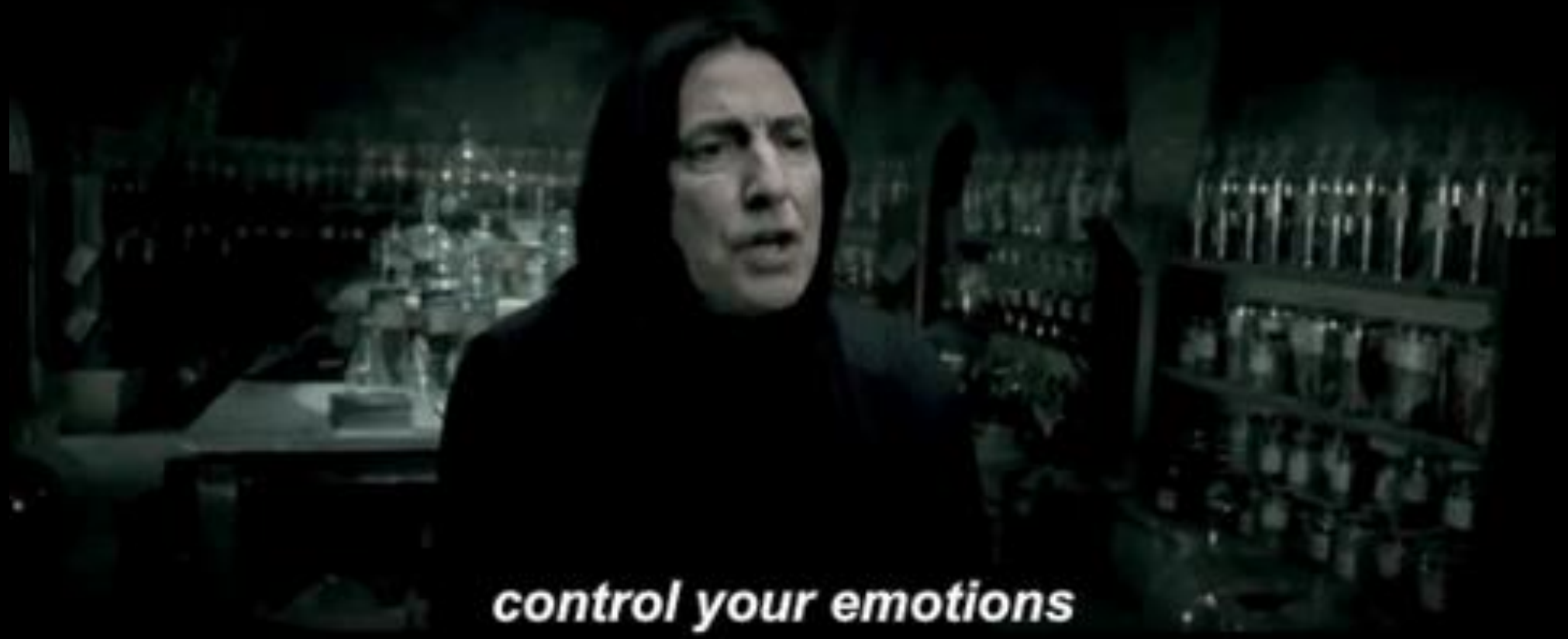
VOLGENDE WEEK

- Interacties
- Dummy-variabelen
- Rapporteren

TOT SLOT

- De slides zijn online beschikbaar (https://elmarjansen.nl/os)
- Een handout / reader met de samenvatting is in de maak
- Deel je opmerkingen / wensen / vragen!
  elmar@elmarjansen.nl

BONUS: CONTROLLEREN VOOR MEDIATIE

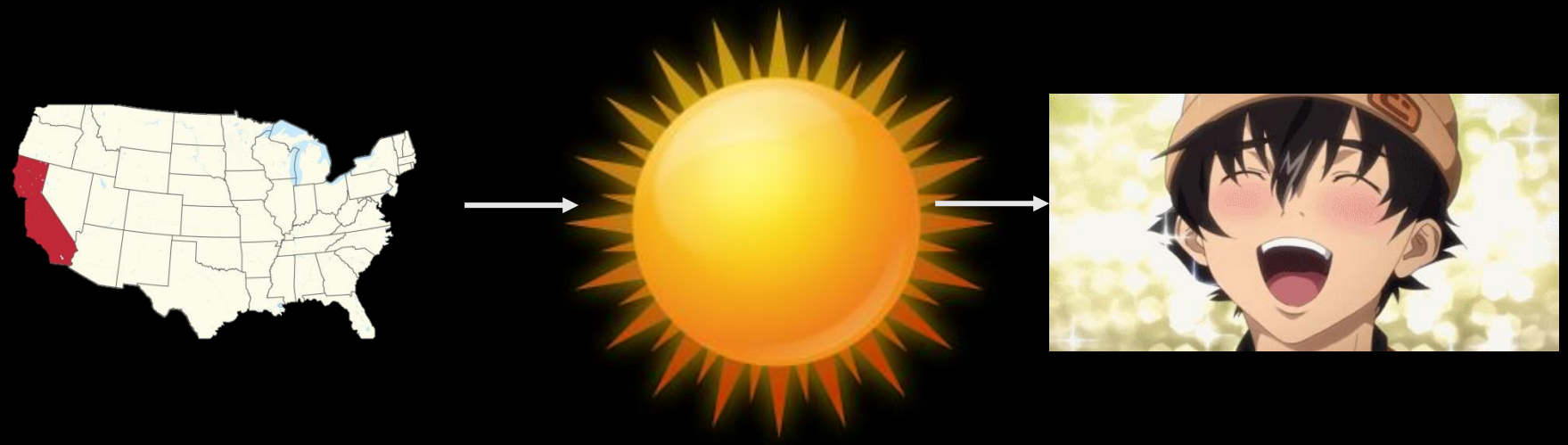


MEDIATIE: VOORBEELD

Tussenstappen

in de

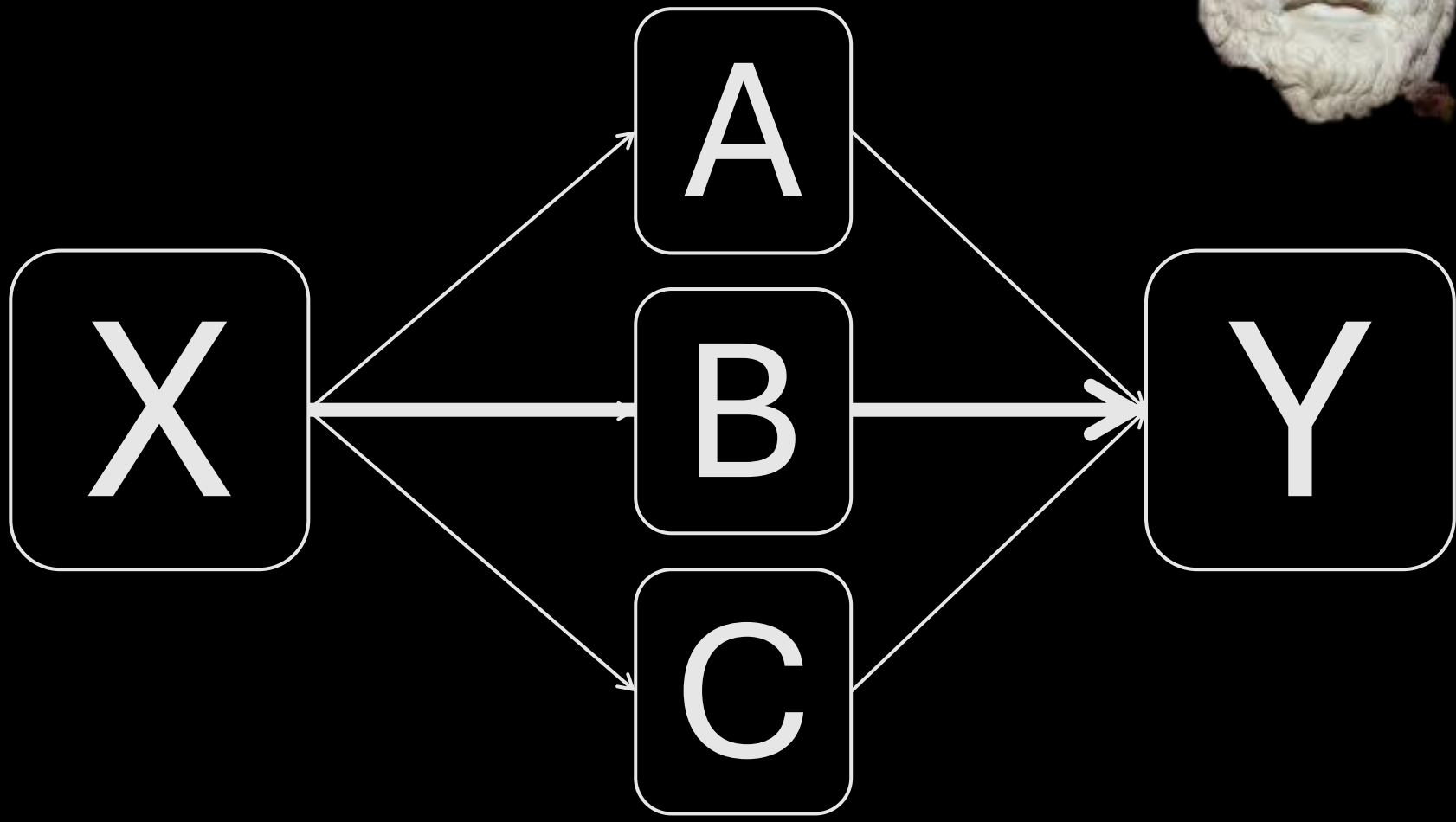
causale keten



Schkade, D. A., & Kahneman, D. (1998). Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction. *Psychological Science*, *9*(5), 340-346. https://doi.org/10.1111/1467-9280.00066

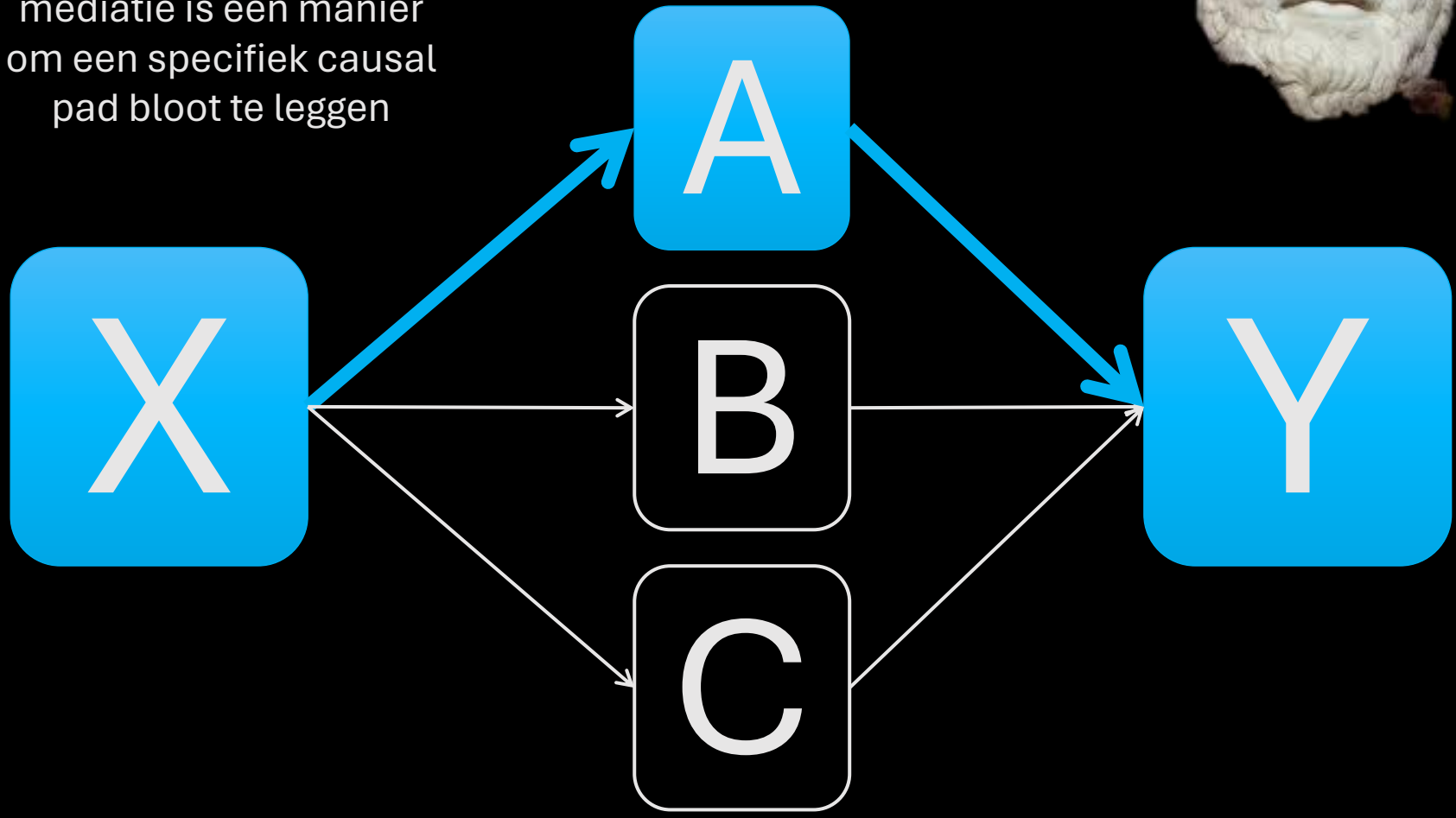
CAUSALE KETENS

CAUSALE KETENS



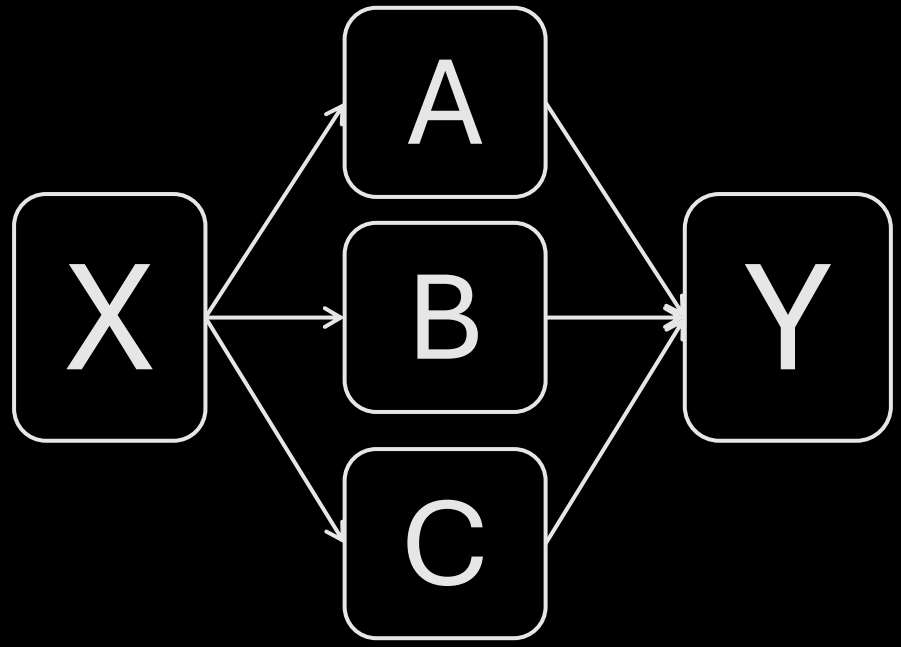
CAUSALE KETENS

Controleren voor mediatie is een manier om een specifiek causal pad bloot te leggen

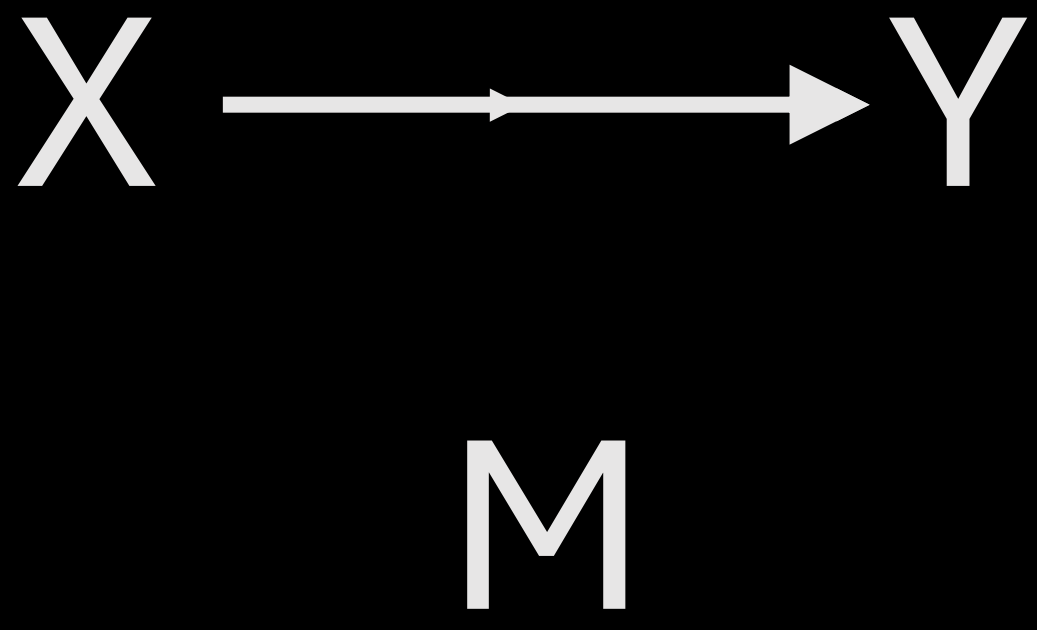


CAUSALE KETENS

Zeggen dat een relatie "gemedieerd is", is op zichzelf niet zo interessant: ieder verband is altijd op te delen in tussenstapjes



MEDIATIE



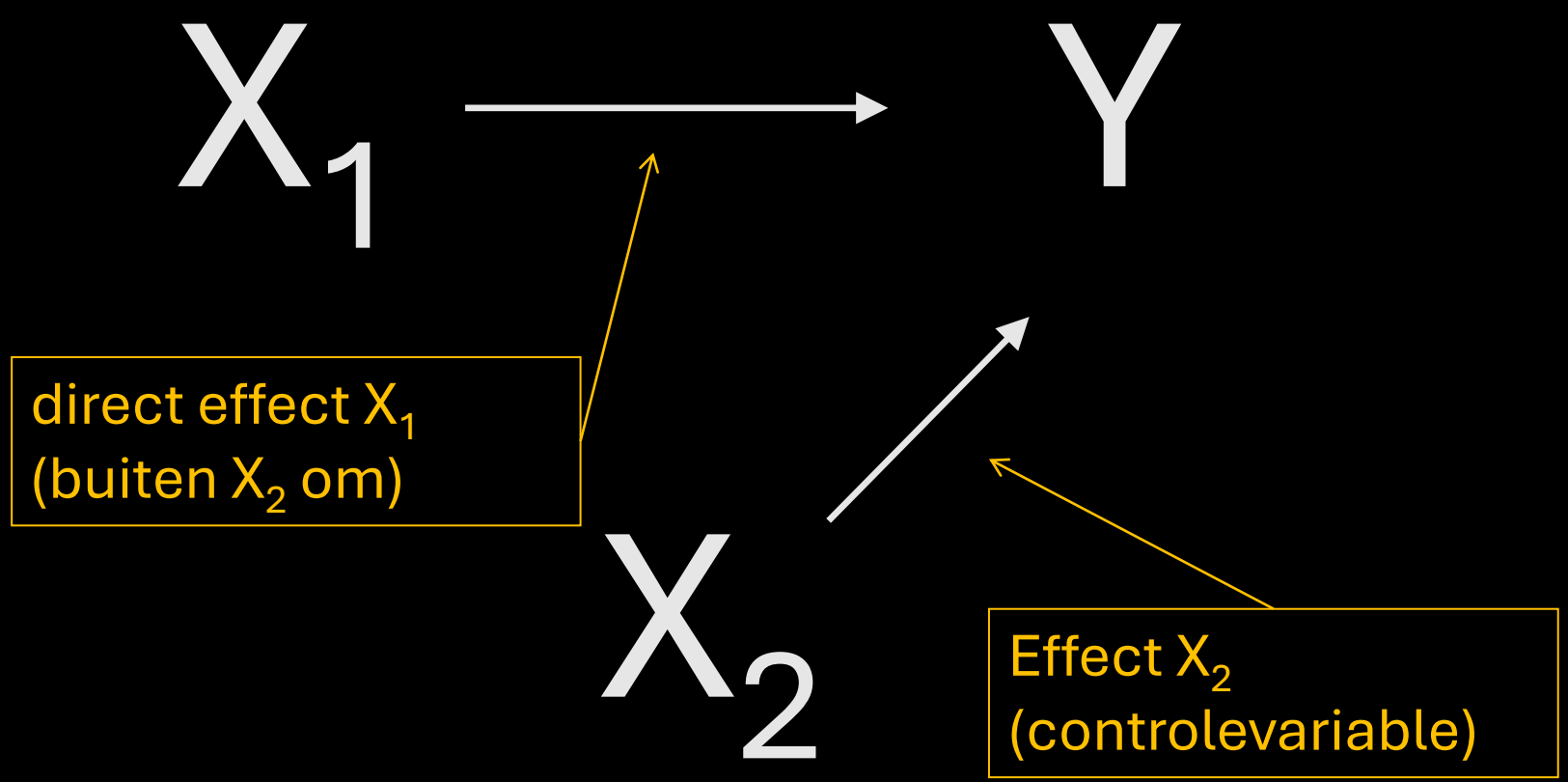
MEDIATIE

+±?

+

+

IN REGRESSIE



REGRESSIE EN CAUSALITEIT

Inkomen

Werkloosheid

???

Geluk

EERSTE MODEL (GEEN CONTROLES)

```
Call:
lm(formula = happy ~ uempla, data = ess10)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2667 -1.2667  0.7333  1.3447  3.3447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.26674    0.01017  714.44   <2e-16 ***
uempla      -0.61145    0.05048  -12.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.93 on 37519 degrees of freedom
  (90 observations deleted due to missingness)
Multiple R-squared:  0.003895,  Adjusted R-squared:  0.003868
F-statistic: 146.7 on 1 and 37519 DF,  p-value: < 2.2e-16
```

Totaal effect van werkloosheid op geluk

Als je werkloos wordt, gaat geluk omlaag met 0.6 (op 10-punts-schaal)

TWEEDE MODEL (CONTROLES)

```
Call:
lm(formula = happy ~ uempla + hinctnta, data = ess10)

Residuals:
    Min      1Q  Median      3Q
-8.1133 -0.9389  0.2355  1.2355

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.369504   0.025002 254.755  < 2e-16 ***
uempla      -0.354982   0.056721  -6.258 3.94e-10 ***
hinctnta     0.174378   0.004116  42.368  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.879 on 29332 degrees of freedom
  (8276 observations deleted due to missingness)
Multiple R-squared:  0.06159,   Adjusted R-squared:  0.06152
F-statistic: 962.5 on 2 and 29332 DF,  p-value: < 2.2e-16
```

Effect werkloosheid op geluk
voor zover niet door inkomensverlies

Als je werkloos wordt en je inkomen blijft gelijk, gaat geluk omlaag met 0.35 (op 10-punts-schaal)

MEDIATIE

Gebruikt om het causale pad te onderzoeken

Methode:

Step 1: Denk na. Bedenk wat potentieel interessante medierende variabelen zijn

Step 2: Draai de regressie met controlevariebelen als extra *onafhankelijke variabelen*

Het gevonden effect van de oorspronkelijke onafhankelijke variabele is nu het deel van het effect dat *niet* via de controlevariabelen loopt.
Oftewel: “het effect van je onafhankelijke variabele bij constante [control variable]”.

CONTROLEREN

Let op: statistisch geen verschil tussen controleren voor *mediatie* of *schijnverband*.

Dit is een keuze die jij moet maken op basis van ***theorie***.

Let op: statistisch geen verschil tussen je *controlevariable(n)* en je primaire onafhankelijke variabele

Dit is alleen een conceptueel verschil.

CONTROLEREN

Voor schijnverband

Oplossing voor probleem

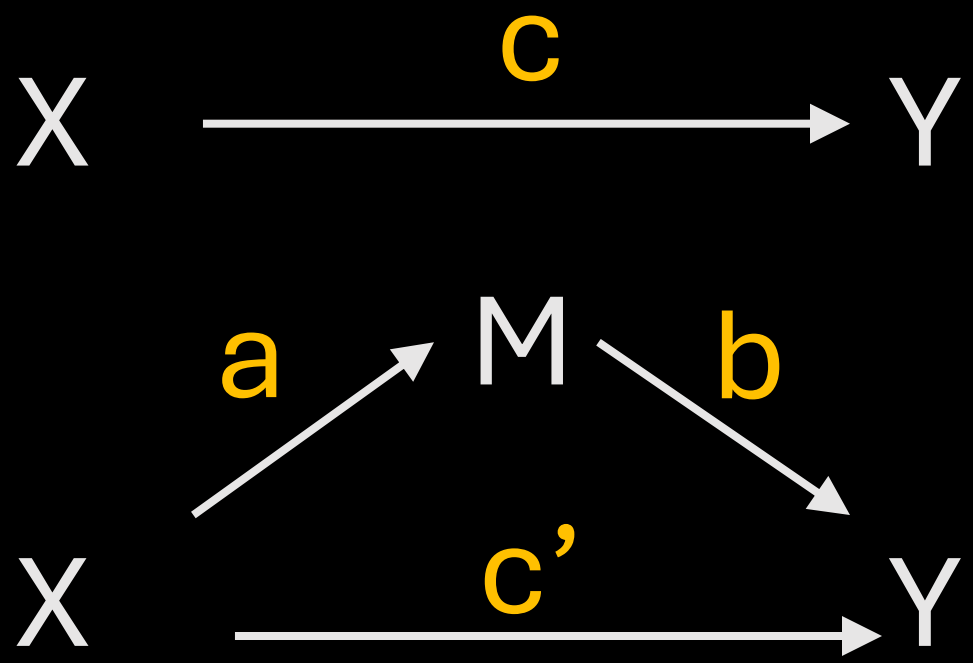
Kies controle op theoretische gronds: wat kan de *confounder zijn*?

Voor mediatie

Approach to find out more

Kies controle op theoretische gronds: wat kan de *tussenliggende variabele zijn*?

EXTR



totaal effect = direct effect + indirect effect

c = c’ + a*b